МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«ПЕНЗЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

С. В. Рындина

АНАЛИЗ ДАННЫХ: ПОСТРОЕНИЕ ОТЧЕТОВ С YANDEX DATALENS И GLARUS BI

Учебно-методическое пособие

ПЕНЗА 2025

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Пензенский государственный университет» (ПГУ)

С. В. Рындина

Анализ данных: построение отчетов с Yandex DataLens и Glarus BI

Учебно-методическое пособие

Пенза
Издательство ПГУ
2025

УДК 004.738.5(075)+339(075)
Р95

Р е ц е н з е н т
кандидат экономических наук, доцент,
руководитель сектора методологии учета и оформления
банковских операций ПАО Банк «Кузнецкий»
О. Б. Волошина

Рындина, Светлана Валентиновна.

Р95 Анализ данных: построение отчетов с Yandex DataLens и Glarus BI : учеб.-метод. пособие / С. В. Рындина. – Пенза : Изд-во ПГУ, 2025. – 88 с.

Рассматриваются возможности систем Business Intelligence: Yandex DataLens и Glarus BI для анализа данных. Приведены задания для лабораторных работ по разведочному анализу данных. Материал пособия соответствует программе дисциплины «Анализ данных» и может быть использован при изучении дисциплин «Бизнес-аналитика на основе больших данных», «Интеллектуальные информационные системы», а также при написании выпускной работы бакалавра.

Издание подготовлено на кафедре «Цифровая экономика» и предназначено для обучающихся по направлению подготовки 38.03.05 «Бизнес-информатика».

УДК 004.738.5(075)+339(075)

© Пензенский государственный
университет, 2025

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 4

1. РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ В GLARUS BI 5
   1.1. Система Business Intelligence (Glarus BI) 5
   1.2. Импорт данных. Типы данных 6
   1.3. Построение запросов 11
   1.4. Построение дашбордов 30
   1.5. Объединение данных и визуализаций 40

2. РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ
В YANDEX DATALENS 55
   2.1. Yandex DataLens – система Business Intelligence 55
   2.2. Создание подключения и датасета 58
   2.3. Создание чартов 62
   2.4. Создание дашборда 69
   2.5. Объединение данных 76

Лабораторная работа № 1. Создание отчета
по разведочному анализу данных средствами Glarus BI 85

Лабораторная работа № 2. Создание отчета
по разведочному анализу данных средствами Yandex Datalens 86

СПИСОК ЛИТЕРАТУРЫ 87

ВВЕДЕНИЕ

В соответствии с учебным планом студенты второго курса направления подготовки 38.03.05 «Бизнес-информатика» изучают дисциплину «Анализ данных».

Описательный анализ данных, разведочный анализ данных и проверка статистических гипотез – базовые методы работы с данными.

Целью учебно-методического пособия является рассмотрение возможностей BI-систем Yandex DataLens и Glarus BI для разведочного анализа данных.

Рассматривается построение запросов и дашбордов в Glarus BI, а также создание подключений, датасетов, чартов и дашбордов в Yandex DataLens.

Также приведены задания лабораторных работ по анализу данных в Yandex DataLens и Glarus BI.

1. РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ В GLARUS BI

1.1. Система Business Intelligence (Glarus BI)

Glarus BI – это российская BI-платформа уровня Enterprise для бизнес-аналитики [1]. Сервис позволяет подключаться к различным источникам данных, строить визуализации, собирать дашборды и делиться полученными результатами. За основу разработчики платформы взяли open source BI-систему Metabase. Glarus BI включена в Реестр программного обеспечения Минцифры (Запись в реестре от 30.01.2023 № 16417).

Glarus BI состоит из нескольких сущностей, которые обеспечивают полный цикл работы с данными (рис. 1.1):

– запрос (SQL-запрос);

– дашборд (dashboard, приборная панель) – набор визуализаций по запросам, фильтров, текстовых блоков, ссылок и других элементов;

– коллекция;

– модель.

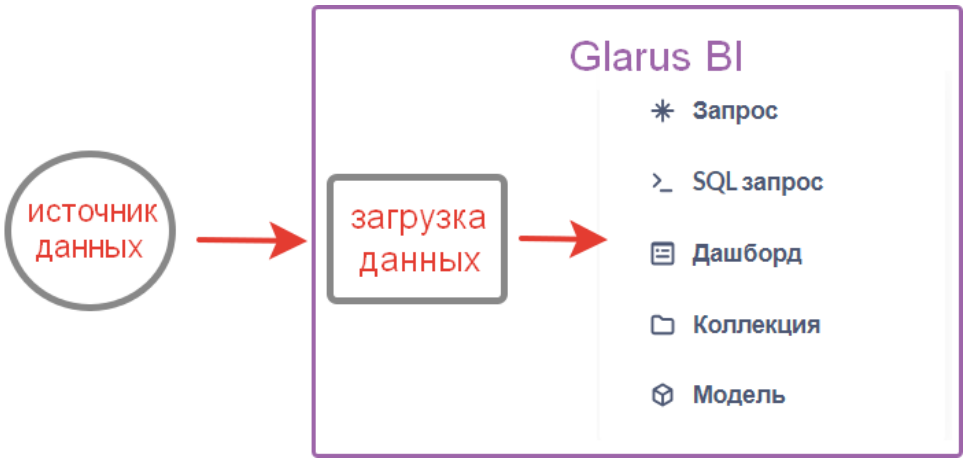


Рис. 1.1. Сущности Glarus BI

Коллекция – папка, в которой хранятся запросы, дашборды, модели. *Модель* – это набор данных или запрос, который можно снабдить метаданными (т.е. описанием отдельных показателей). В настоящем пособии модели рассматриваться не будут.

Платформу можно установить на хостинге, принадлежащем организации, или сразу развернуть облачную виртуальную машину с Glarus BI у таких облачных провайдеров, как VK cloud, MWS или

T1 cloud, но в пособии будет рассматриваться только вариант работы в облаке https://demo.glarus-analytics.ru, доступ к которому можно получить автоматически, заполнив на сайте https://glarus-bi.online регистрационную форму, которая открывается щелчком на кнопке (рис. 1.2).

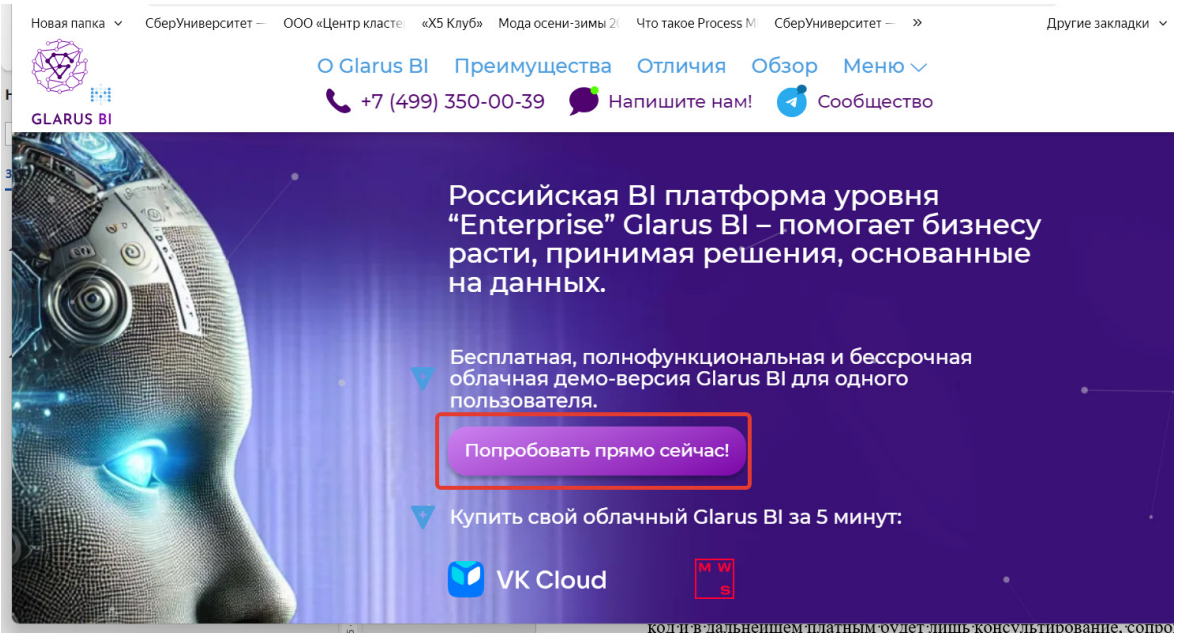


Рис. 1.2. Облачная демоверсия для одного пользователя

Пользовательский доступ для тестирования возможностей платформы по работе с данными предоставляется бесплатно. Роль пользователя не единственная из возможных ролей на платформе. Так, роль администратора позволяет подключать новых пользователей, осуществлять разграничение прав доступа пользователям и распределять ресурсы. Доступ к платформе в роли администратора возможен при коммерческом использовании платформы и по академической программе для учебных организаций.

В коммерческих целях выгоднее произвести установку на собственном хостинге, тем более что компания, разработавшая систему, передает исходный код и в дальнейшем платным будет лишь консультирование, сопровождение и обучение при запросе на эти услуги со стороны заказчика.

Любой проект по анализу данных начинается с данных. У данных есть тип и структура.

1.2. Импорт данных. Типы данных

Выполним разведочный анализ данных для набора данных insurance.csv (доступен по ссылке https://www.kaggle.com/

mirichoi0218/insurance). На ресурсе kaggle.com для скачивания файла с данными необходимо зарегистрироваться.

Файл с именем insurance.csv представляет собой текстовый файл с данными, разделенными запятой. Такое представление данных называется плоскими таблицами. Можно просмотреть содержимое файла с помощью текстового редактора, не поддерживающего форматирование, например редактора Блокнот (рис. 1.3).

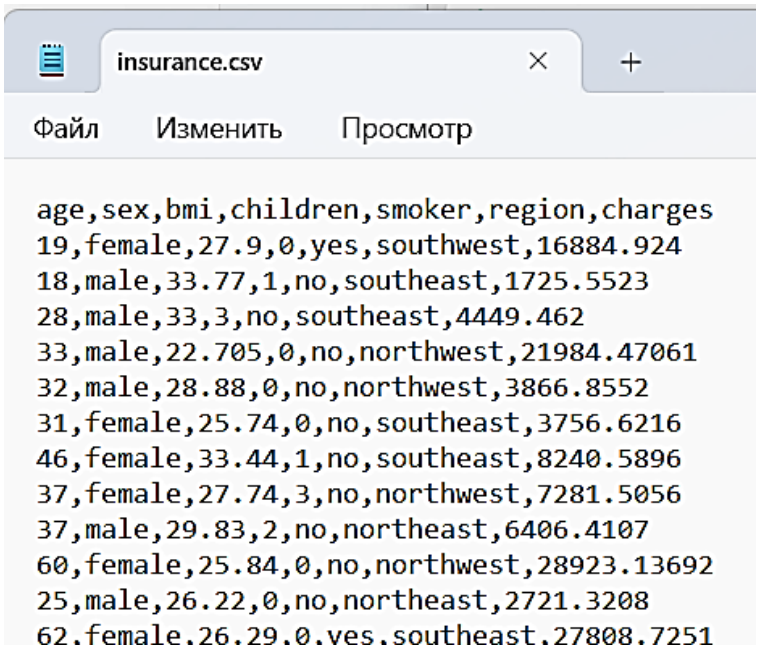


Рис. 1.3. Просмотр данных

В демоверсии Glarus BI требуется определенное расширение файла (.xlsx).

Можно воспользоваться MS Excel, открыв новый файл и выполнив импорт данных из текстового файла (рис. 1.4).

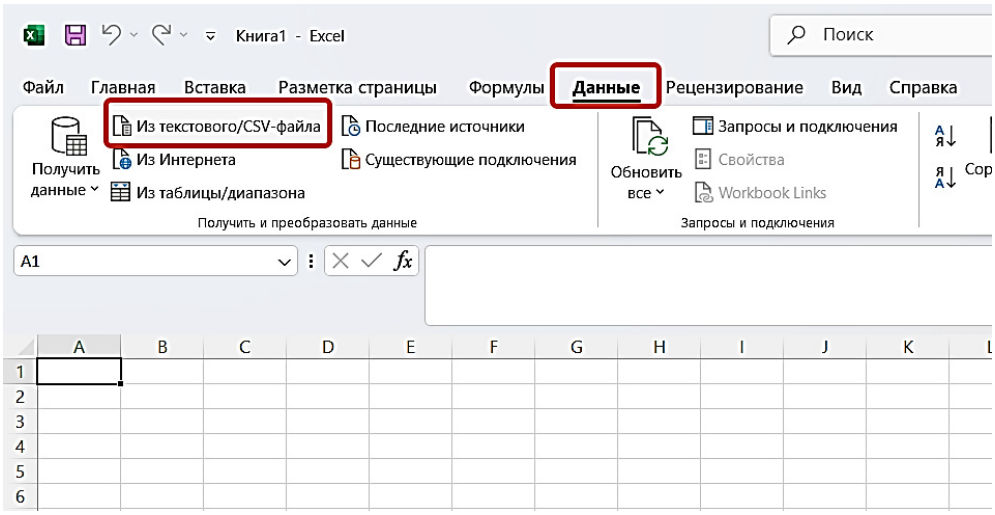
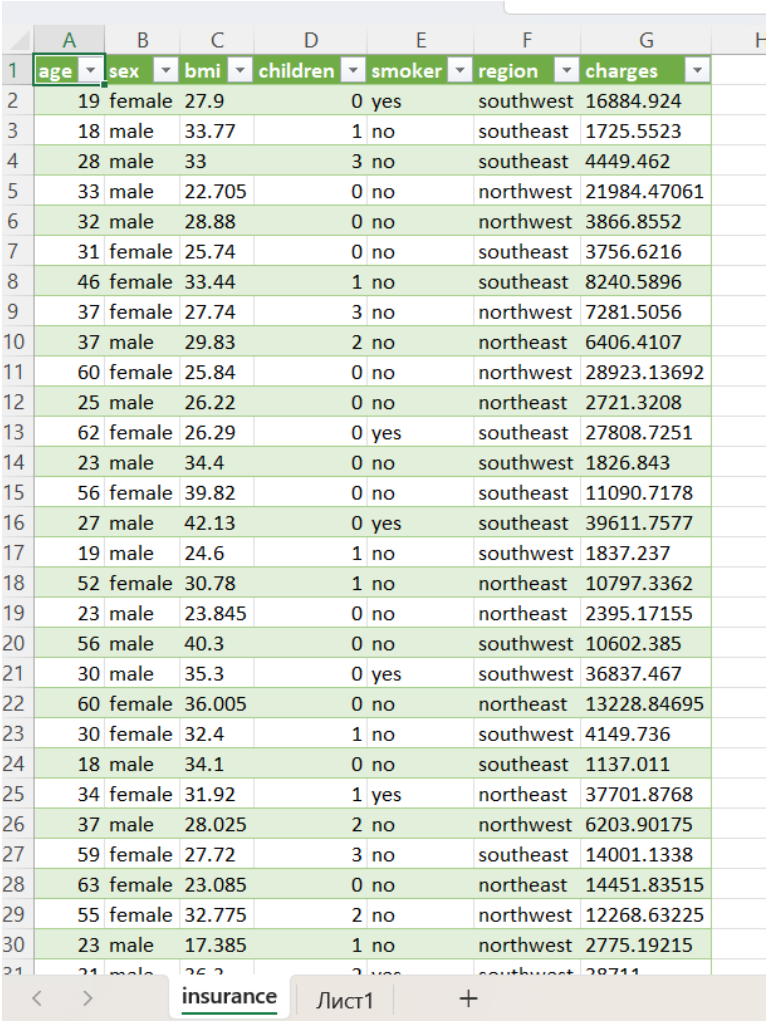


Рис. 1.4. Импорт данных из текстового файла

С помощью Мастера импорта нужно выполнить необходимые настройки и получить данные в таблице (рис. 1.5), сохранить файл в требуемом формате.



	A	B	C	D	E	F	G
1	age	sex	bmi	children	smoker	region	charges
2	19	female	27.9	0	yes	southwest	16884.924
3	18	male	33.77	1	no	southeast	1725.5523
4	28	male	33	3	no	southeast	4449.462
5	33	male	22.705	0	no	northwest	21984.47061
6	32	male	28.88	0	no	northwest	3866.8552
7	31	female	25.74	0	no	southeast	3756.6216
8	46	female	33.44	1	no	southeast	8240.5896
9	37	female	27.74	3	no	northwest	7281.5056
10	37	male	29.83	2	no	northeast	6406.4107
11	60	female	25.84	0	no	northwest	28923.13692
12	25	male	26.22	0	no	northeast	2721.3208
13	62	female	26.29	0	yes	southeast	27808.7251
14	23	male	34.4	0	no	southwest	1826.843
15	56	female	39.82	0	no	southeast	11090.7178
16	27	male	42.13	0	yes	southeast	39611.7577
17	19	male	24.6	1	no	southwest	1837.237
18	52	female	30.78	1	no	northeast	10797.3362
19	23	male	23.845	0	no	northeast	2395.17155
20	56	male	40.3	0	no	southwest	10602.385
21	30	male	35.3	0	yes	southwest	36837.467
22	60	female	36.005	0	no	northeast	13228.84695
23	30	female	32.4	1	no	southwest	4149.736
24	18	male	34.1	0	no	southeast	1137.011
25	34	female	31.92	1	yes	northeast	37701.8768
26	37	male	28.025	2	no	northwest	6203.90175
27	59	female	27.72	3	no	southeast	14001.1338
28	63	female	23.085	0	no	northeast	14451.83515
29	55	female	32.775	2	no	northwest	12268.63225
30	23	male	17.385	1	no	northwest	2775.19215

Рис. 1.5. Импорт из текстового файла в книгу Excel

Можно воспользоваться онлайн-конвертором и преобразовать файл с расширением .csv в файл с расширением .xlsx (некоторые конверторы требовательны к размеру файла и не будут выполнять конвертацию для очень объемных наборов данных).

Бургерное меню (рис. 1.6) предоставляет доступ к основным сущностям платформы: Ваша личная коллекция – это Запросы, Дашборды и Модели доступные только пользователю соответствующего аккаунта.

Перемещение элементов в Общую коллекцию позволяет открыть к ним доступ и другим пользователям. Если элемент является составным, то в общую коллекцию необходимо вместе с ним переместить и связанные элементы.

Загрузим файл insurance.xlsx на платформу (рис. 1.6).

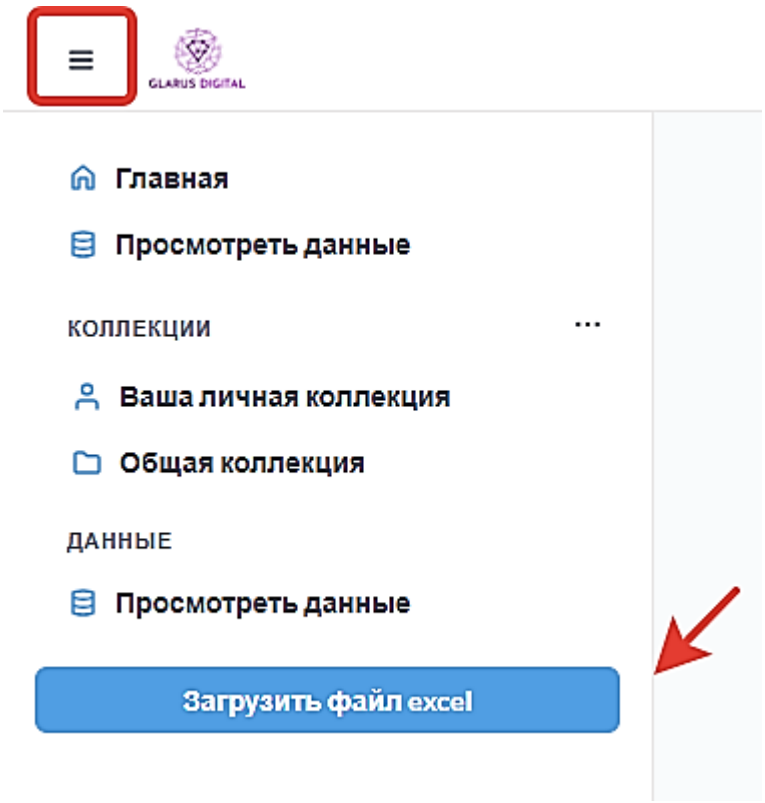


Рис. 1.6. Загрузка файла на платформу

Загруженный файл будет размещен в СУБД ClickHouse_Test, к которой можно получить доступ, выбрав опцию Просмотреть данные (рис. 1.7).

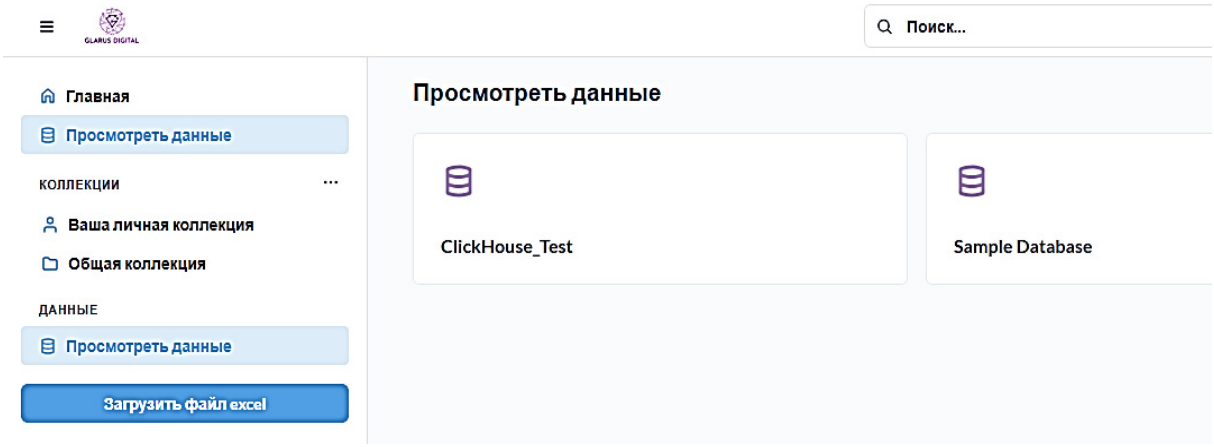


Рис. 1.7. Просмотр данных

Название таблицы с данными в базе данных составлено из имени файла insurance и имени листа с данными Worksheet. При наведении курсора на имя таблицы становится доступно пиктографическое меню, предоставляющее опции сканирования, просмотра сведений о таблице и удаления (рис. 1.8).

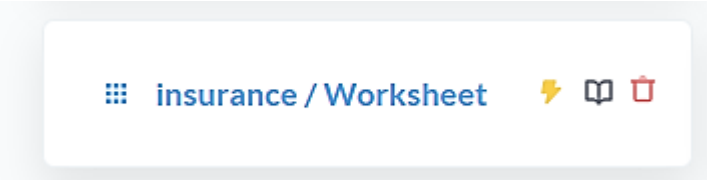


Рис. 1.8. Пиктографическое меню для действий с таблицей

Просмотрим сведения о таблице, проконтролируем корректность распознавания типов данных (рис. 1.9).

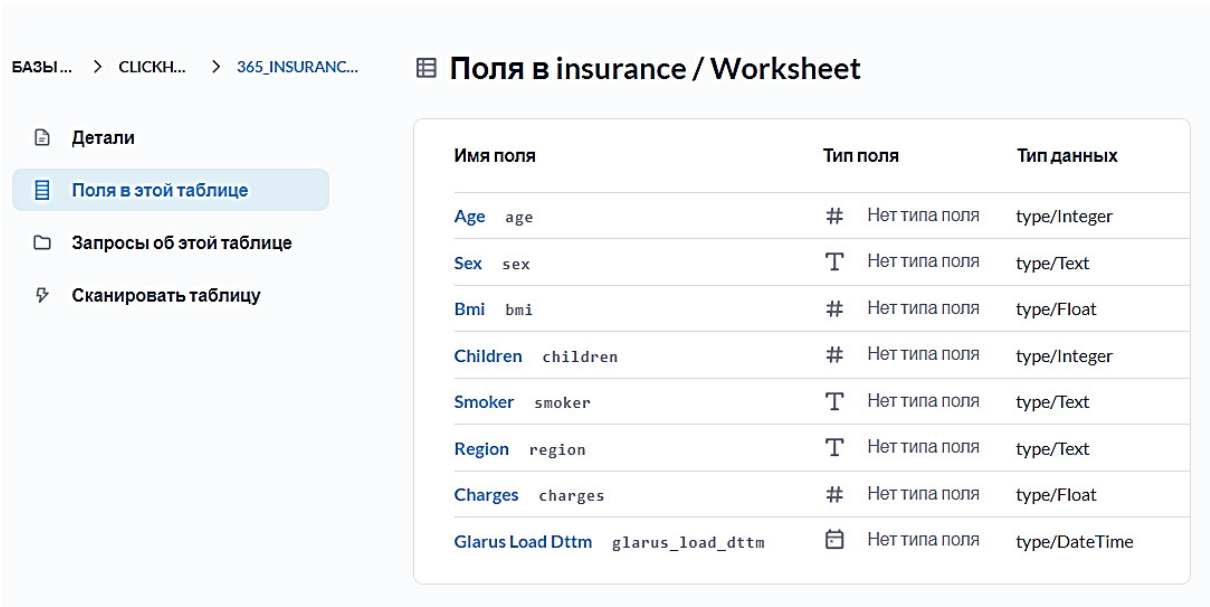


Рис. 1.9. Просмотр типов полей данных для загруженного набора

Описание переменных набора:

– age: возраст основного бенефициара;

– sex: пол застрахованного;

– bmi: индекс массы тела;

– children: число детей, охваченных медицинским страхованием / число иждивенцев;

– smoker: курит ли застрахованный;

– region: жилой район получателя страховых выплат (северо-восток, юго-восток, юго-запад, северо-запад).

– charges: индивидуальные медицинские расходы, оплачиваемые страховкой.

Типы данных, которые есть в наличии в данном наборе:

– числовые данные, имеют маркер #:

• integer – целочисленные данные;

• float – вещественные (десятичные дроби);

– категориальные данные, имеют маркер T:
  • text – строковые данные;
– данные даты и времени, имеют маркер :
  • DateTime – временные данные.

Поле Qlarus Load Dttm с типом Date time в таблицу добавлено платформой (это момент загрузки данных).

1.3. Построение запросов

Разведочный анализ заключается в построении визуализаций, которые выявляют связи между данными.

Построим визуализацию, с помощью которой можно установить, есть ли связь между индексом массы тела респондента и его медицинскими расходами.

Создадим запрос на основе данных, воспользовавшись кнопкой добавления новой сущности (рис. 1.10).

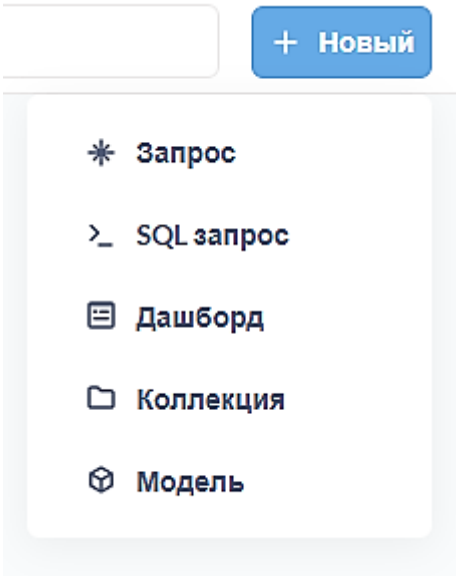


Рис. 1.10. Создание запроса

Выберем начальные данные (рис. 1.11), которые можно просмотреть с помощью стрелки справа в поле имени набора, она же появится, если добавить условие фильтрации или настройки для агрегации (суммирования). К данным можно присоединить дополнительный набор с помощью кнопки .

Построим визуализацию для данных исходной таблицы, выбрав щелчком по кнопке визуализации под таблицей тип Разброс (рис. 1.12).

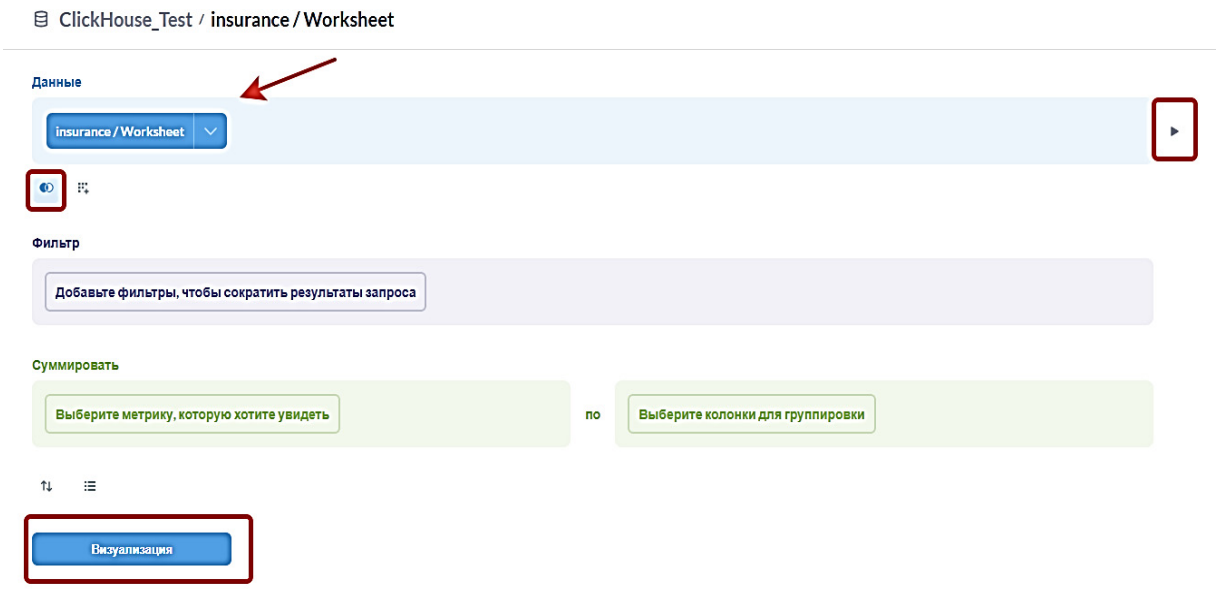


Рис. 1.11. Настройки для запроса

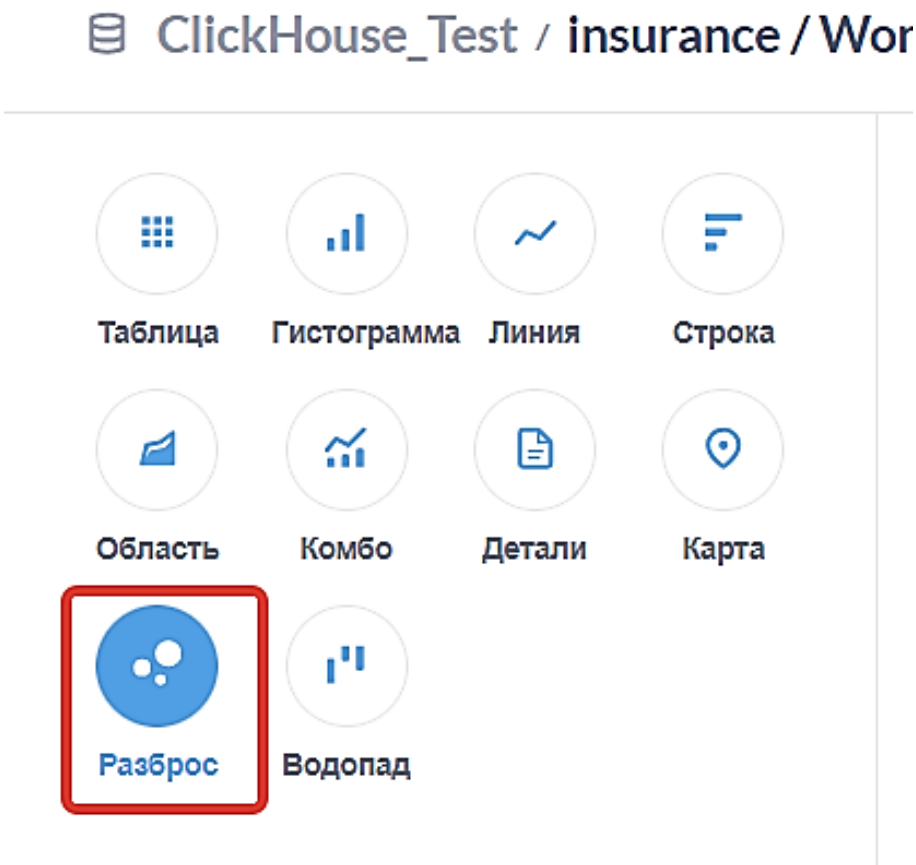


Рис. 1.12. Выбор типа визуализации

В настройках визуализации (рис. 1.13) выберем для оси абсцисс показатель bmi (индекс массы тела), а для оси ординат показатель charges (медицинские расходы).

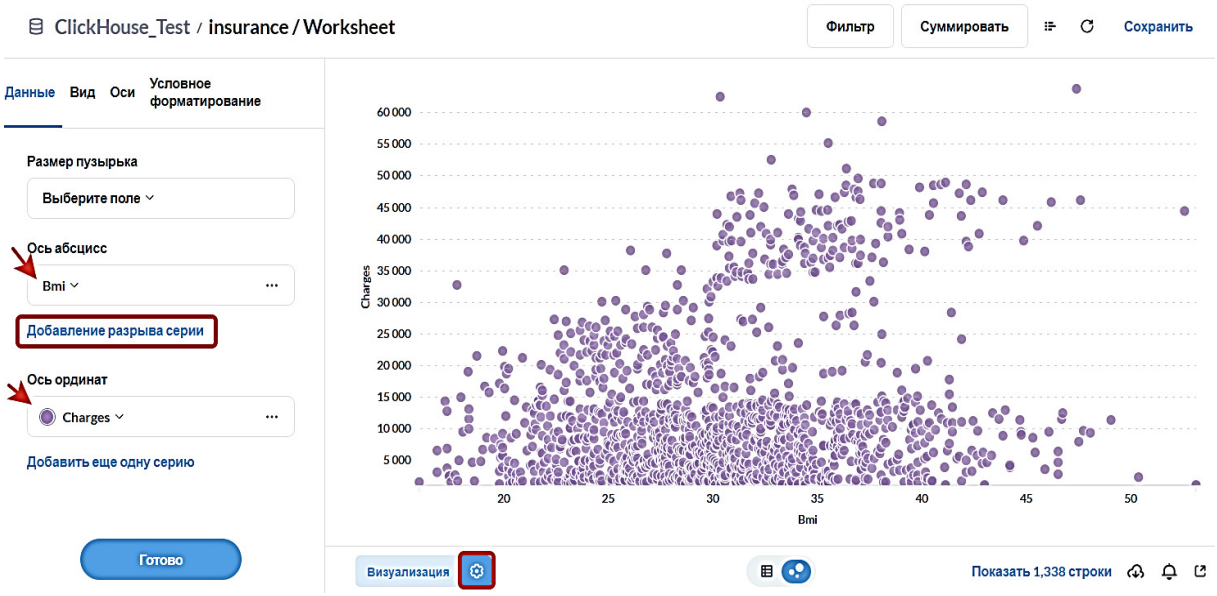


Рис. 1.13. Настройки для визуализации

Построенную визуализацию можно сделать более информативной, добавив разрыв серии (рис. 1.14). Сохраним запрос с названием Диаграмма разброса расходов от веса.

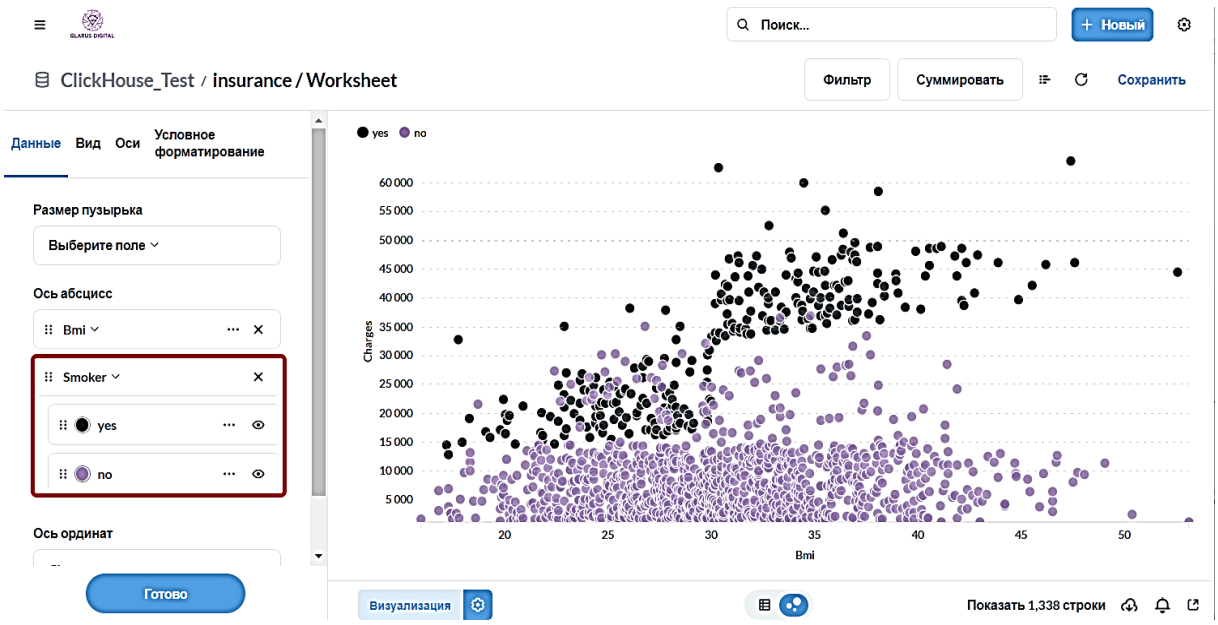


Рис. 1.14. Диаграмма рассеяния с выделением цветом тех, у кого есть вредная привычка

Сохраним запрос с названием Диаграмма разброса расходов от веса (рис. 1.15). Название исходного набора данных оставим в имени.

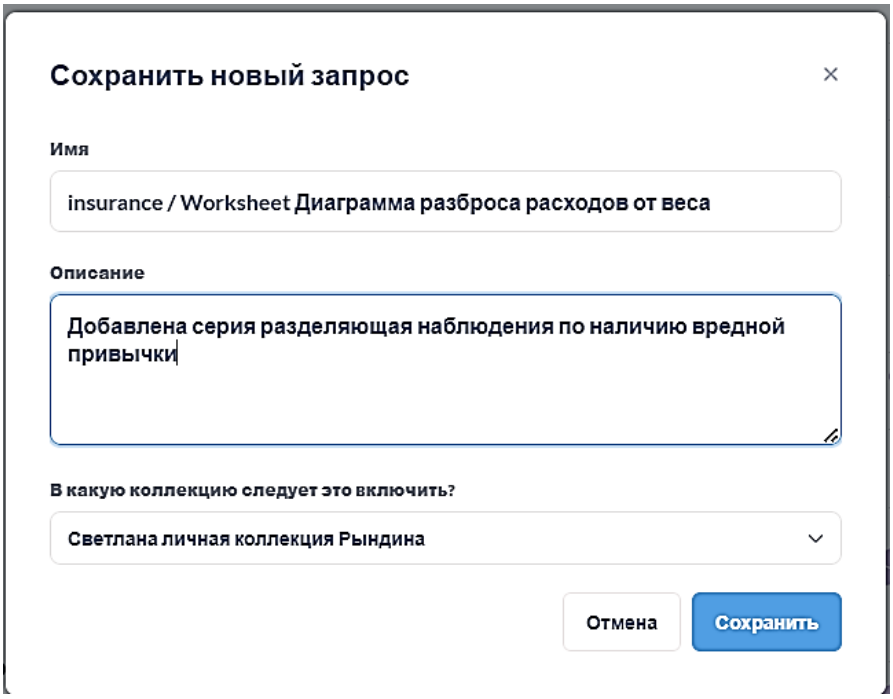


Рис. 1.15. Сохранение запроса

Как видим, расходы зависят и от лишнего веса, и от наличия вредной привычки. Не все обладатели лишнего веса имеют значительные медицинские расходы, однако с увеличением веса вероятность роста расходов повышается. Если потребуется отредактировать сохраненный запрос (рис. 1.16), то кнопки Визуализация позволяют перейти к выбору типа визуализации и ее настройкам. А редактор запросов позволяет изменить конфигурацию запроса, удалить / добавить / изменить фильтр и агрегацию.

Проанализируем сбалансированность набора данных по показателю smoker. Новый запрос можно создать, сразу выбрав исходные данные. Откроем набор данных, щелкнув по его имени в папке ClickHouse_Test, и вызовем редактор запросов (рис. 1.17).

В редакторе (рис. 1.18) выберем для агрегации метрику Количество показателя smoker (рис. 1.19).

В качестве визуализации выберем гистограмму (рис. 1.20) и сохраним запрос с тем названием, которое предложит платформа.

Построим на основе этого запроса новый запрос, определив распространенность вредной привычки в зависимости от района. Для этого продублируем запрос (рис. 1.21, 1.22).

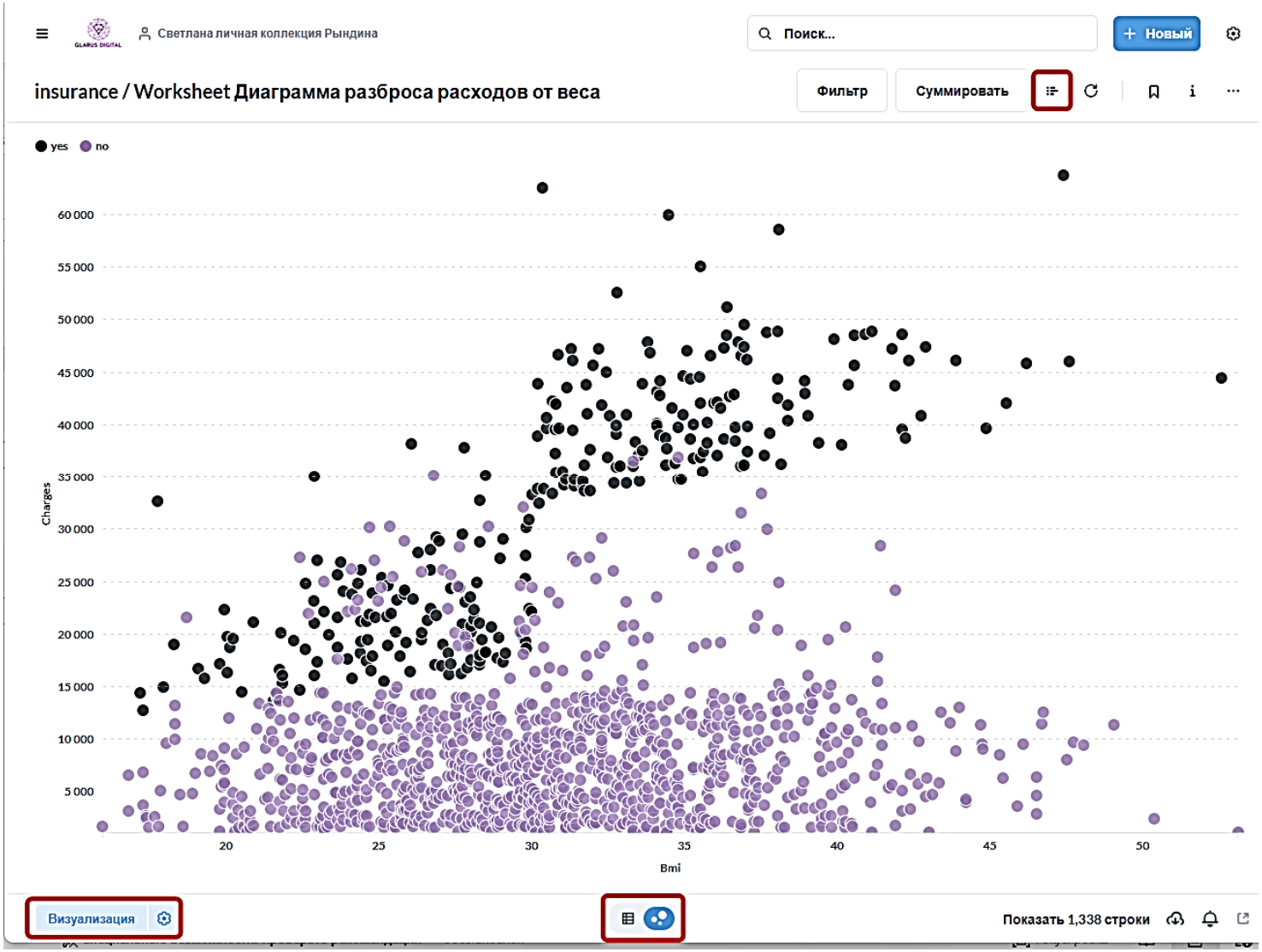


Рис. 1.16. Возможности редактирования запроса

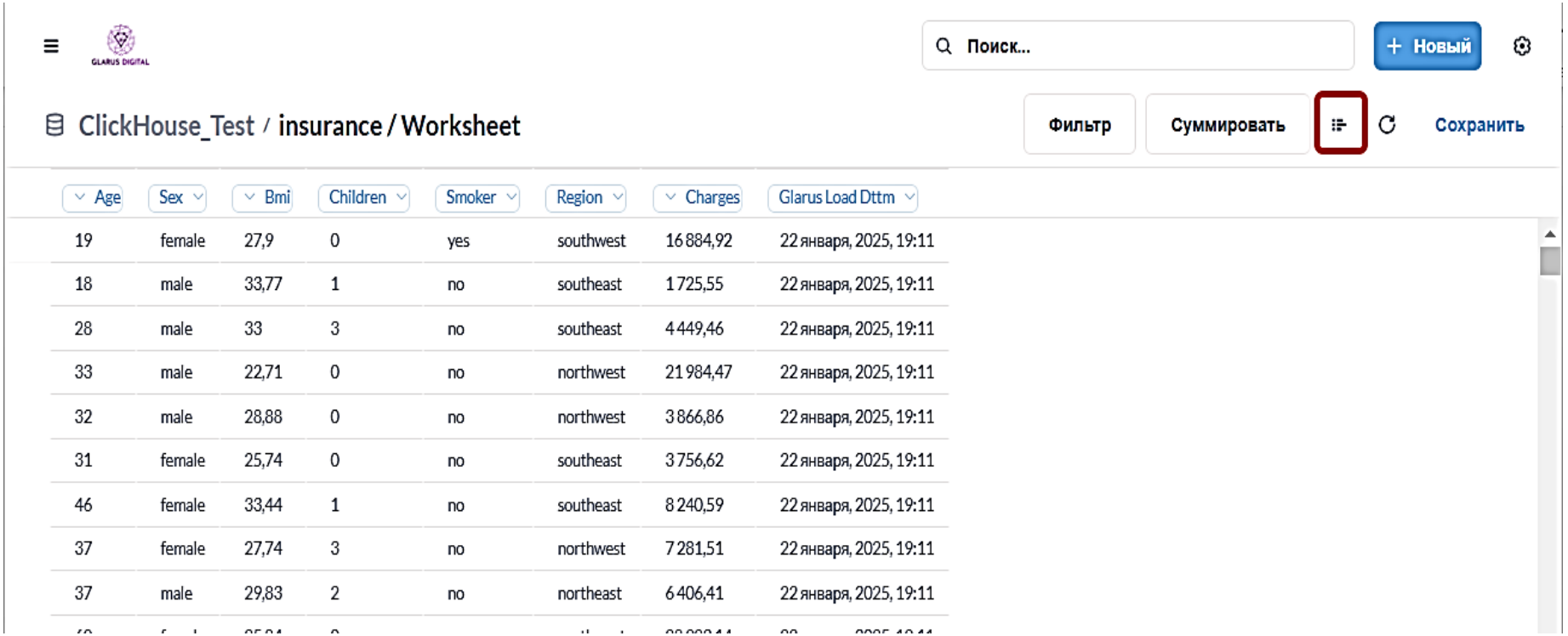


Рис. 1.17. Вызов редактора запросов

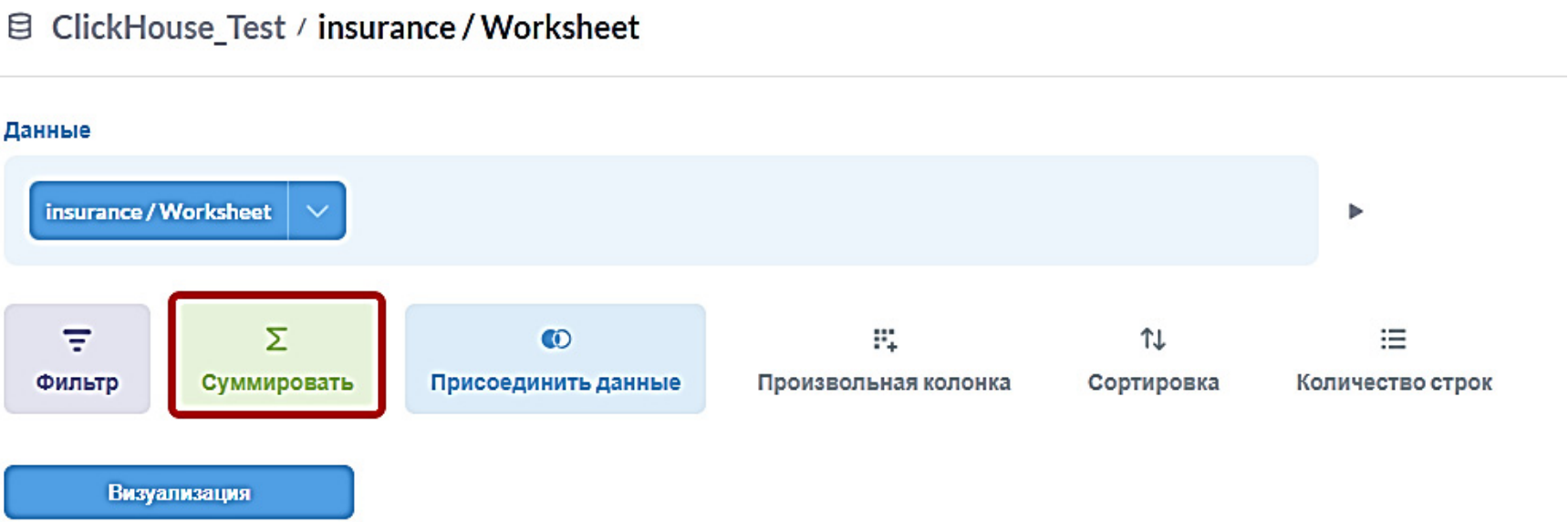


Рис. 1.18. Переход к настройкам агрегации

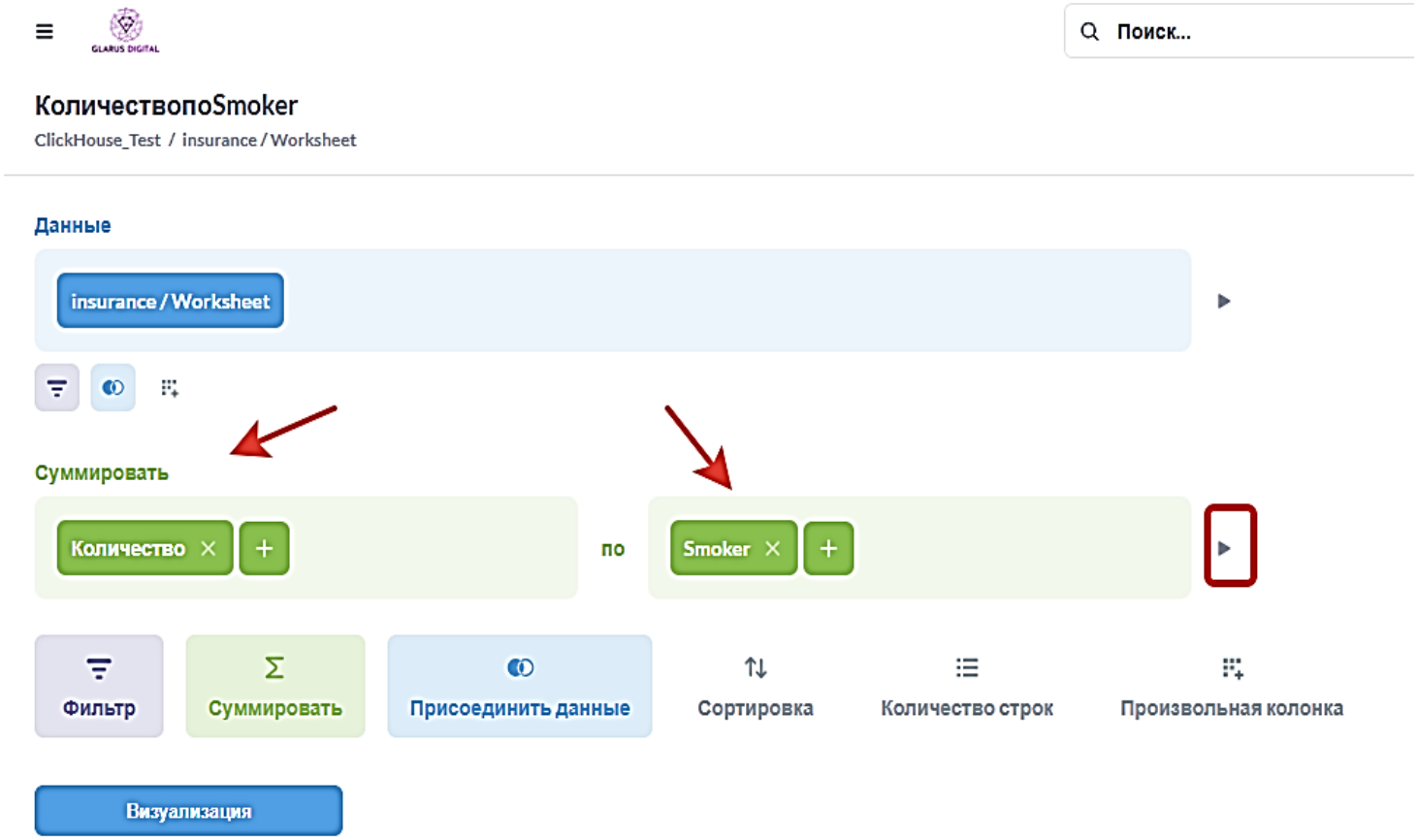


Рис. 1.19. Настройки запроса

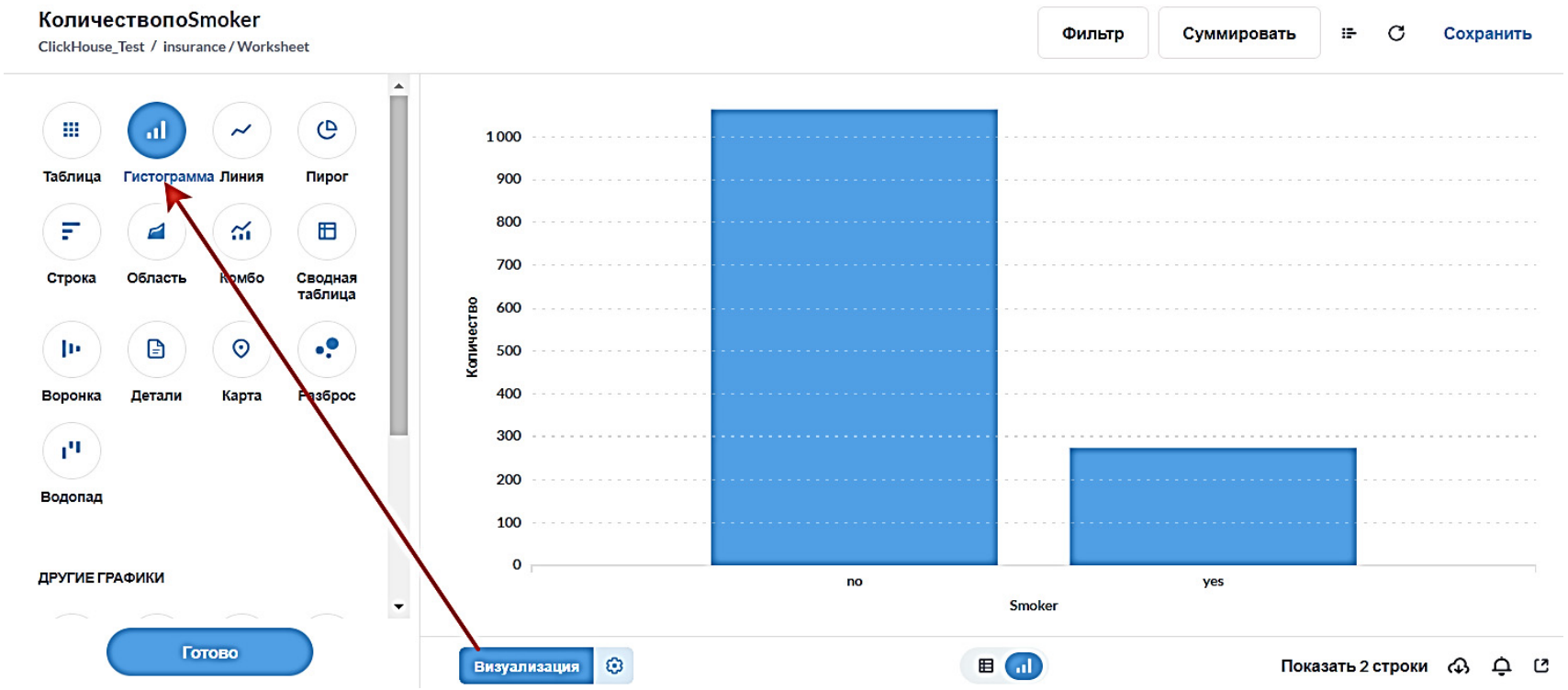


Рис. 1.20. Распространенность вредной привычки в наборе данных

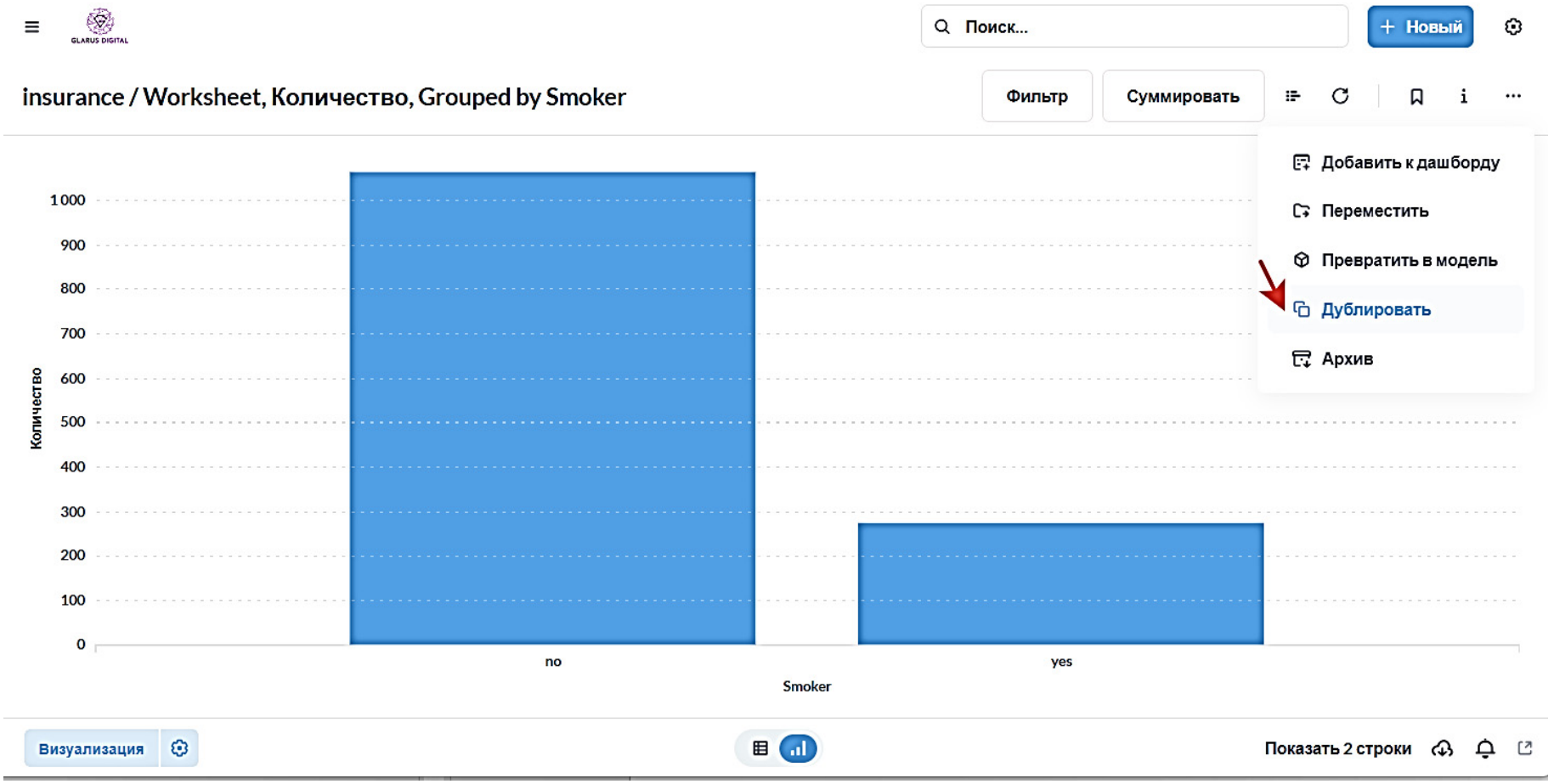


Рис. 1.21. Дублирование запроса

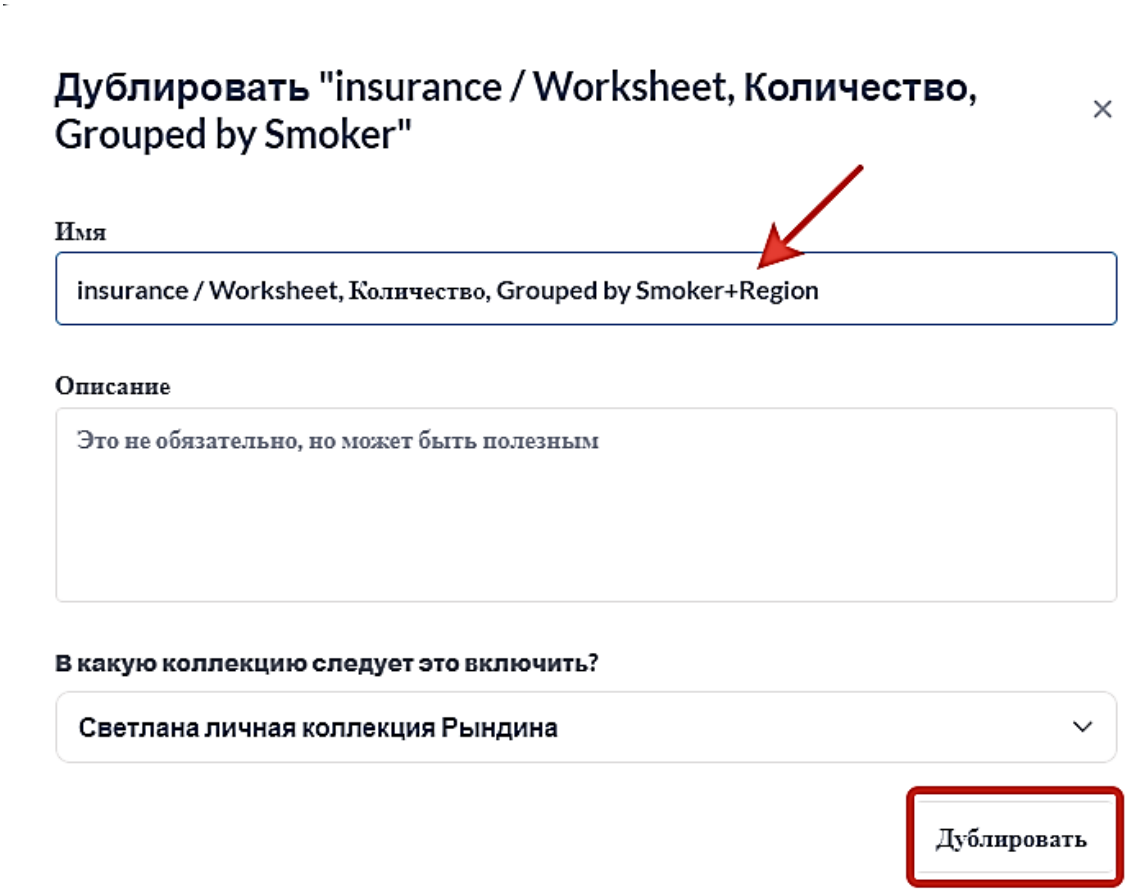


Рис. 1.22. Изменение имени запроса

Перейдем в редактор запроса (рис. 1.23) с помощью кнопки пиктографического меню и добавим в качестве группирующего показателя region (рис. 1.24)

Перейдем к визуализации и в ее настройках изменим показатель для оси абсцисс и выберем опцию Добавление разрыва серии (рис. 1.25). В качестве показателя для разрыва серии выберем smoker, дополнительно скорректируем вид полученной визуализации (рис. 1.26). Сохраним запрос (рис. 1.27).

Проделаем аналогичные манипуляции с ранее построенной диаграммой рассеяния (см. рис. 1.16): перейдем к этому запросу, создадим его копию (рис. 1.28), поменяем показатель для оси абсцисс (вместо bmi выберем age). Сохраним запрос (рис. 1.29).

Создадим визуализацию, анализирующую сбалансированность набора данных по показателю sex (рис. 1.30).

Перейдем к построению дашборда.

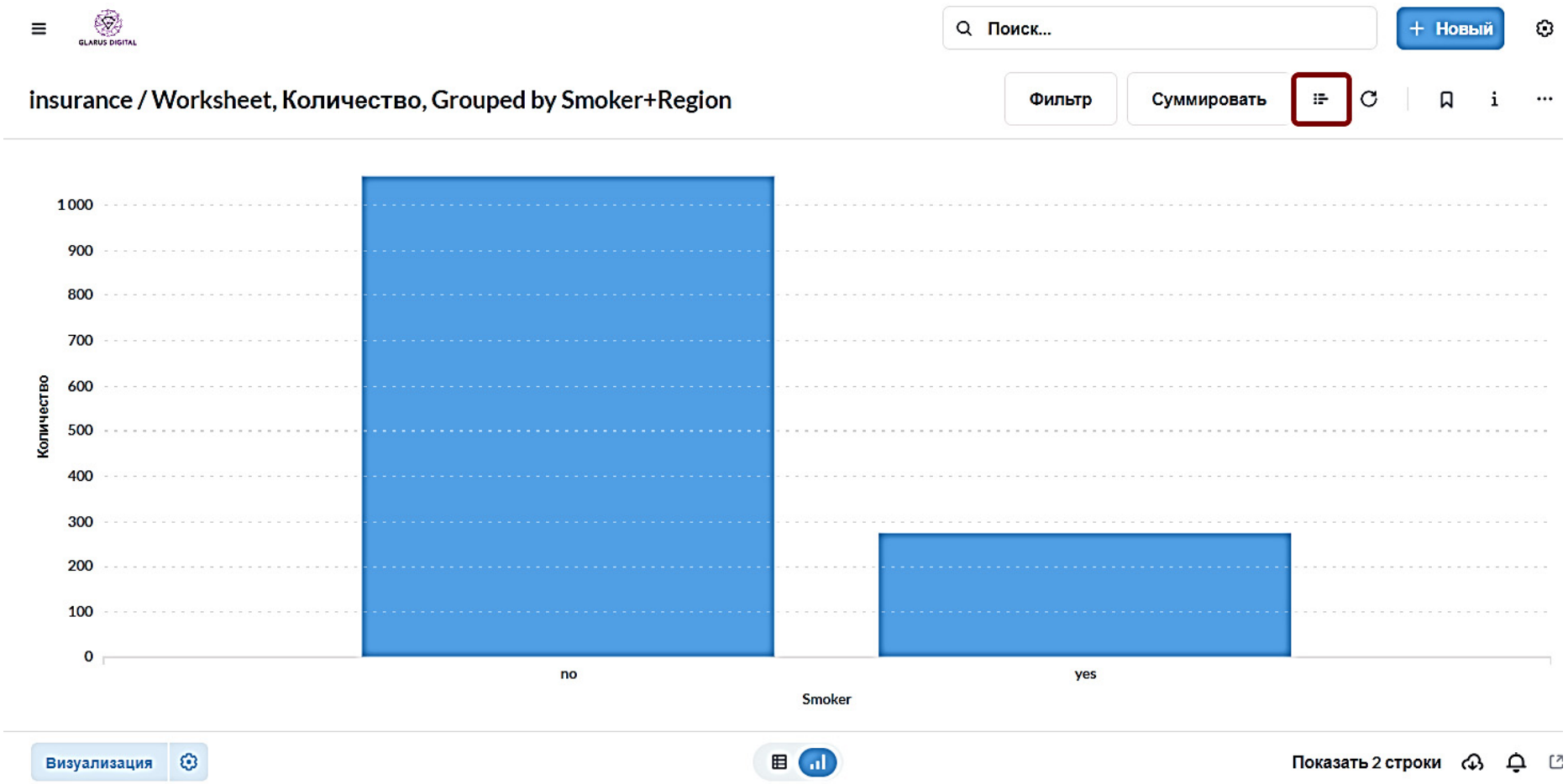


Рис. 1.23. Переход в редактор запросов

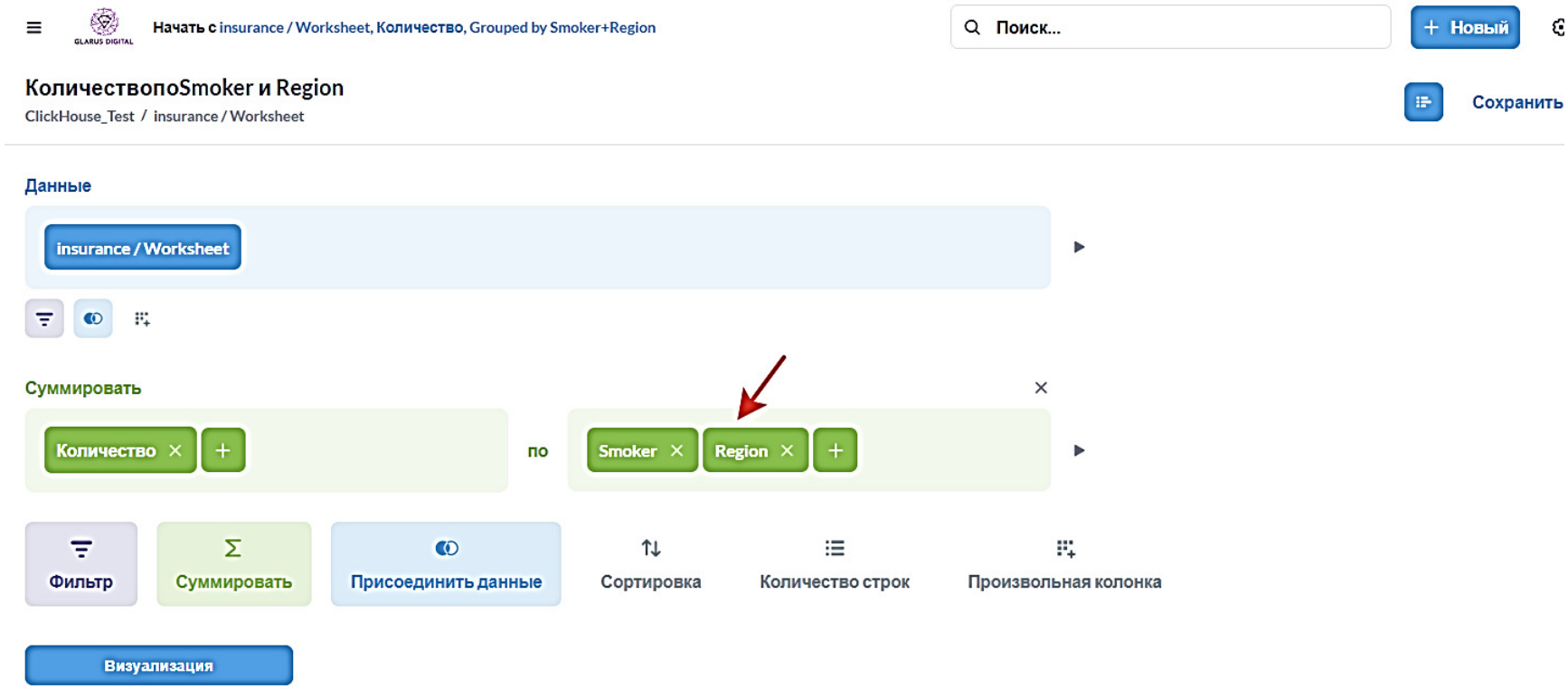


Рис. 1.24. Добавление показателя для группировки наблюдений

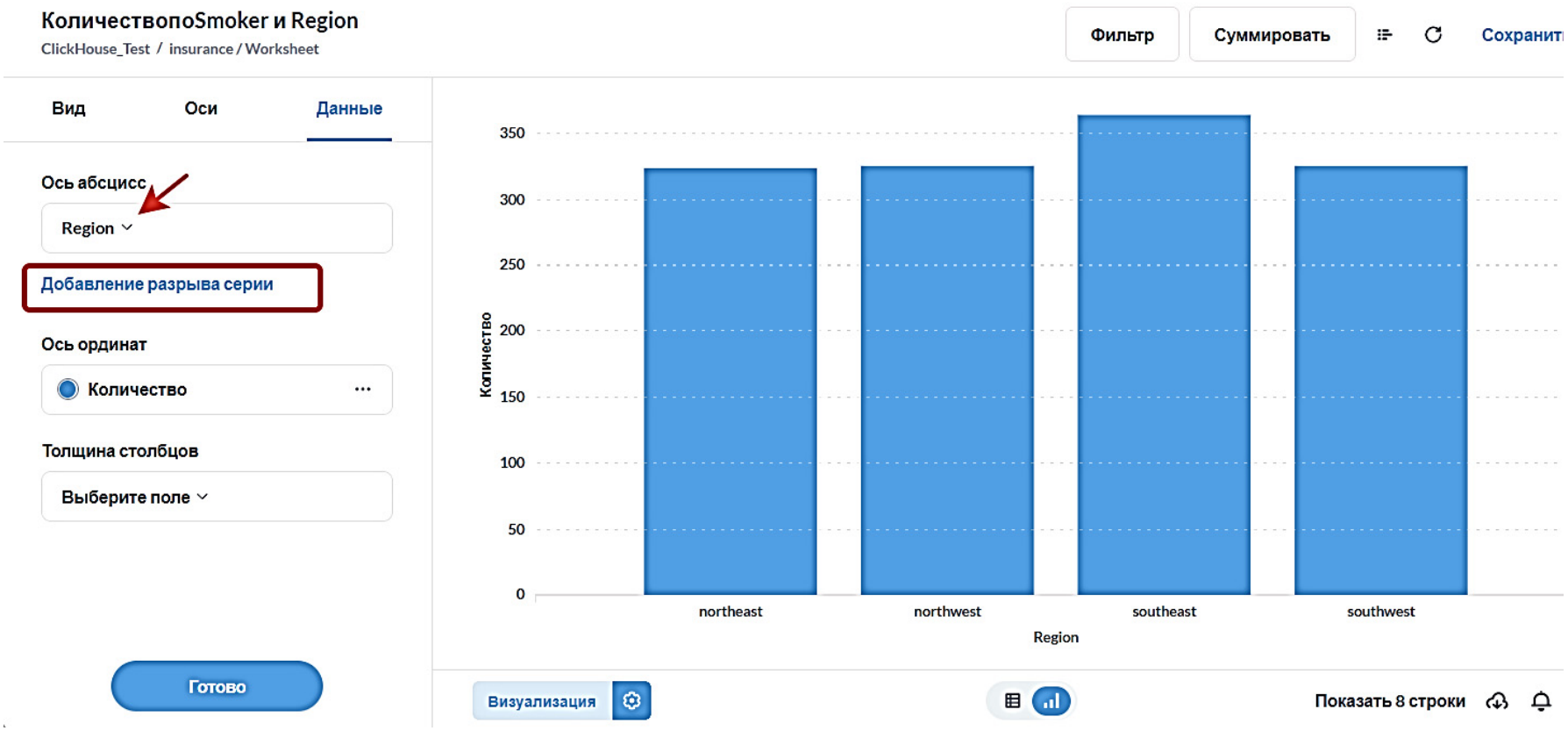


Рис. 1.25. Изменение настроек для визуализации (данные)

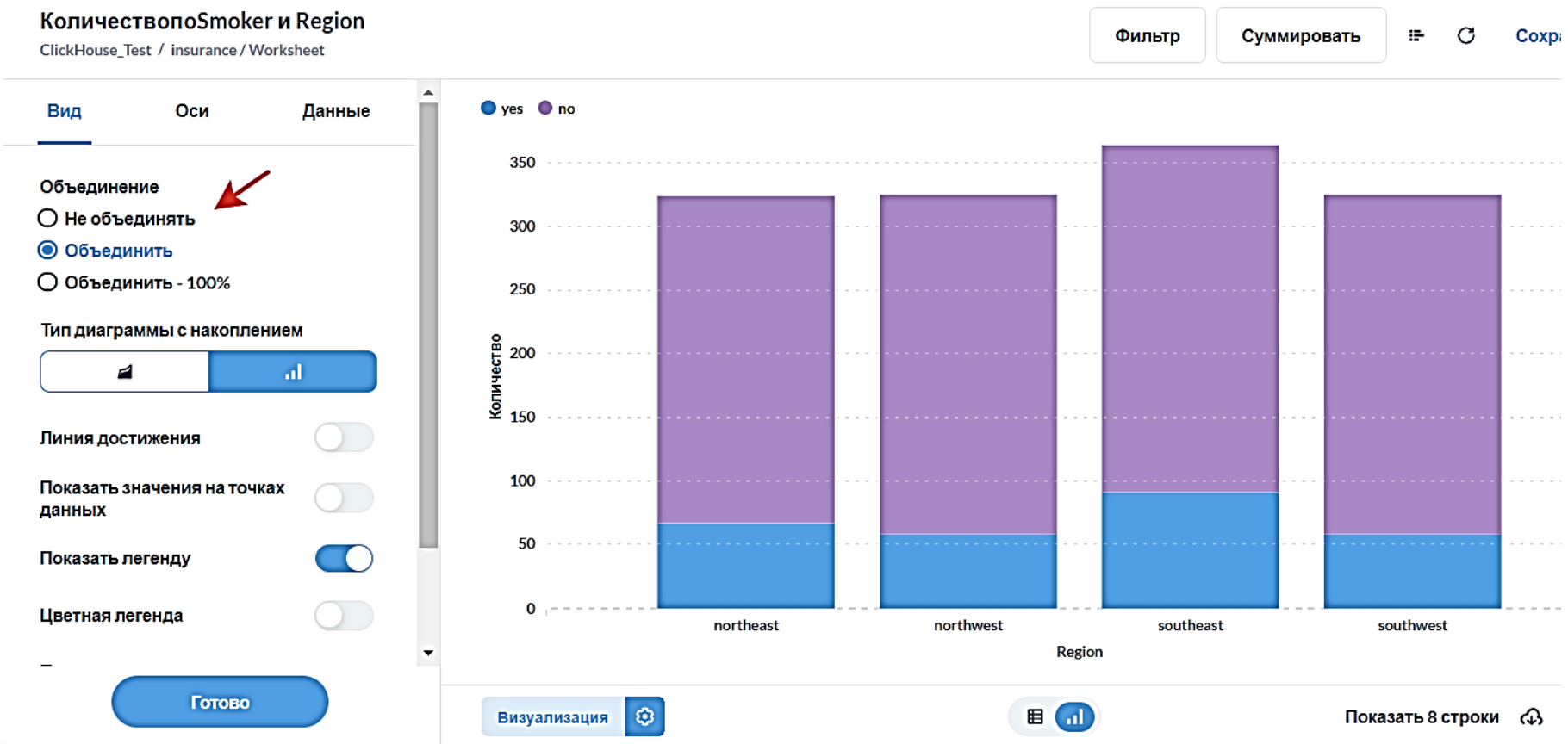


Рис. 1.26. Изменение настроек вида визуализации

Рис. 1.27. Итоговый вид визуализации

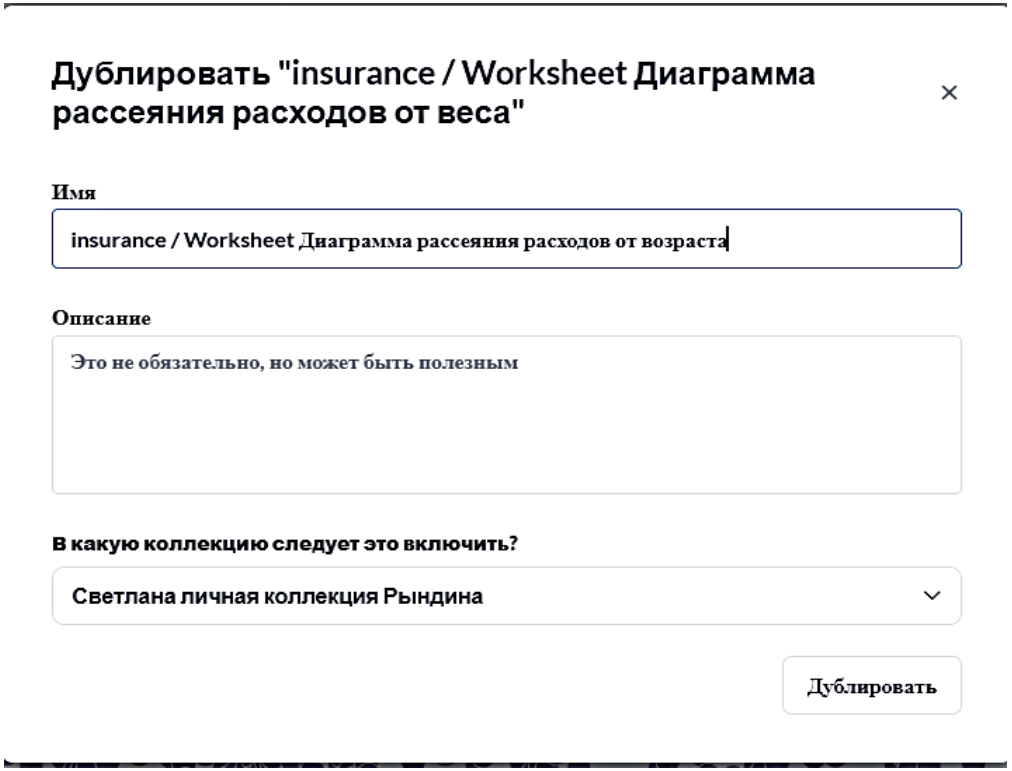


Рис. 1.28. Изменение имени запроса для копии

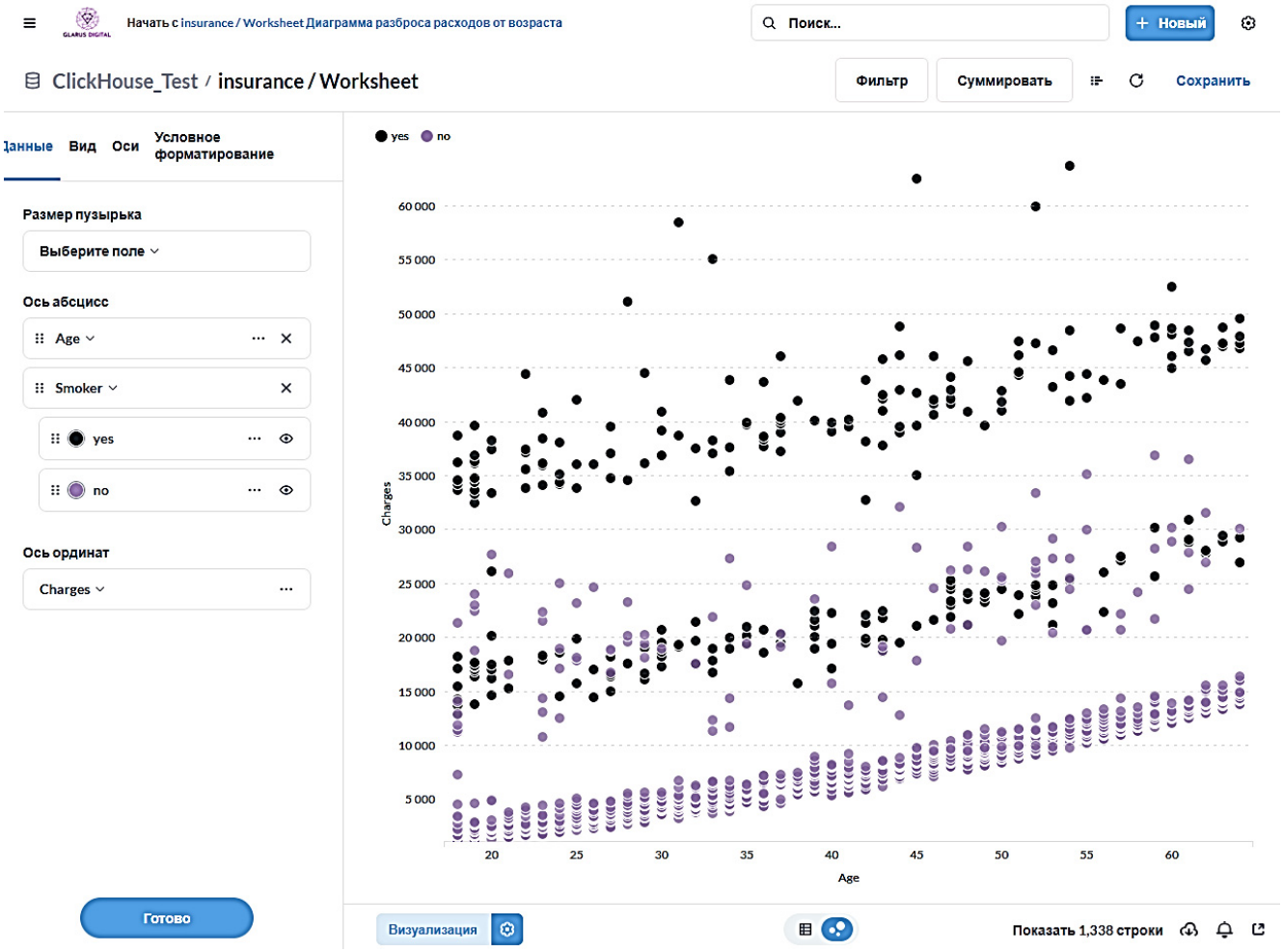


Рис. 1.29. Сохранение визуализации

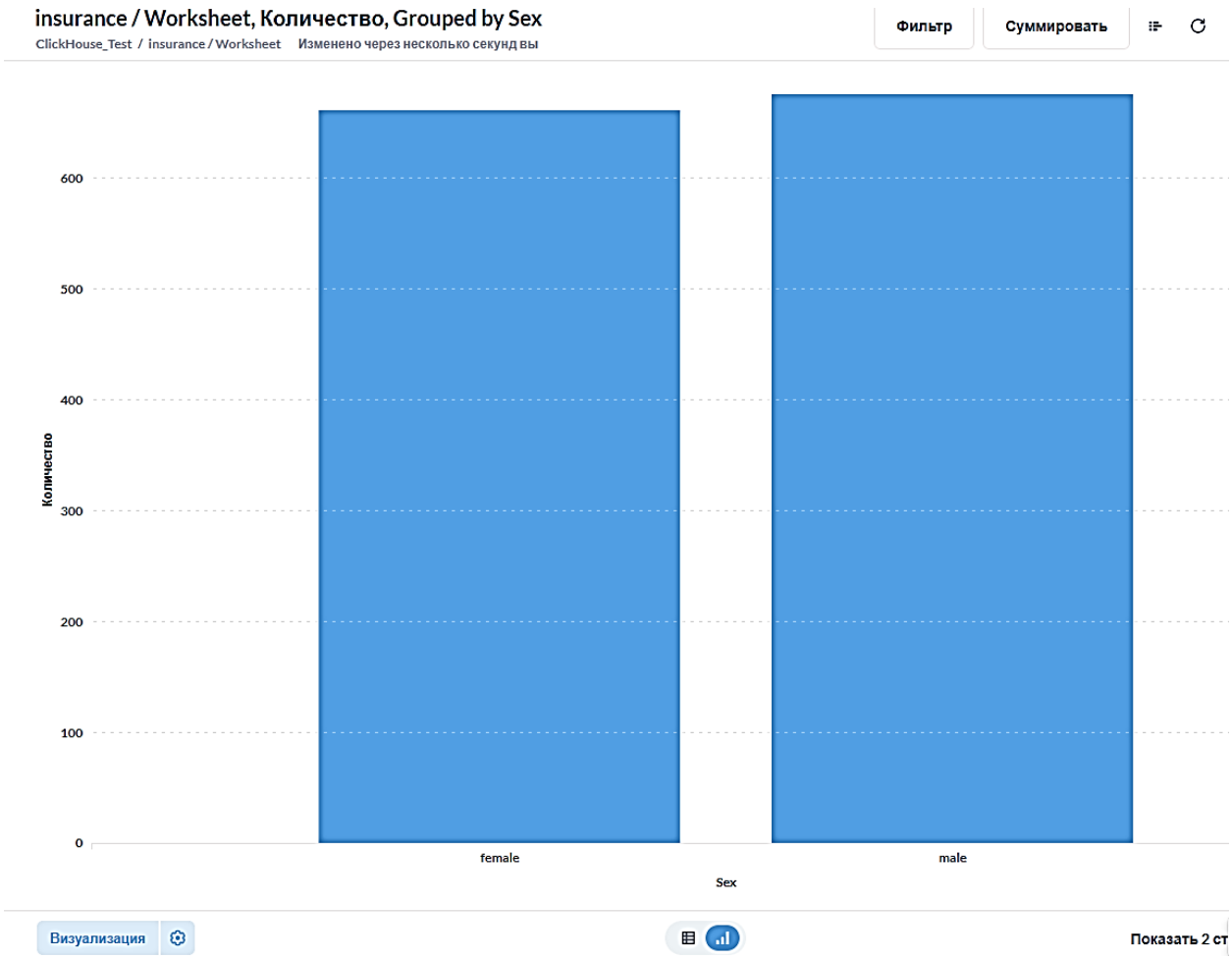


Рис. 1.30. Количество наблюдений с разными значениями показателя sex

1.4. Построение дашбордов

Создадим дашборд с именем Медицинское страхование (рис. 1.31).

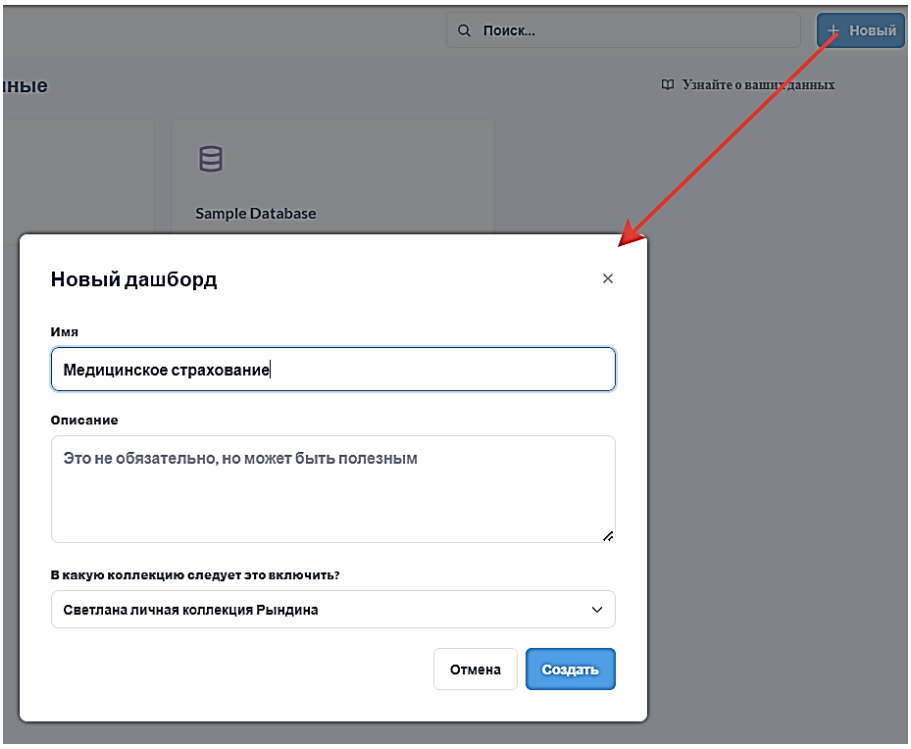


Рис. 1.31. Создание дашборда

Справа вверху в режиме редактирования дашборда имеется пиктографическое меню. Кнопка «+» открывает справа панель с запросами, которые можно добавить на дашборд (рис. 1.32).

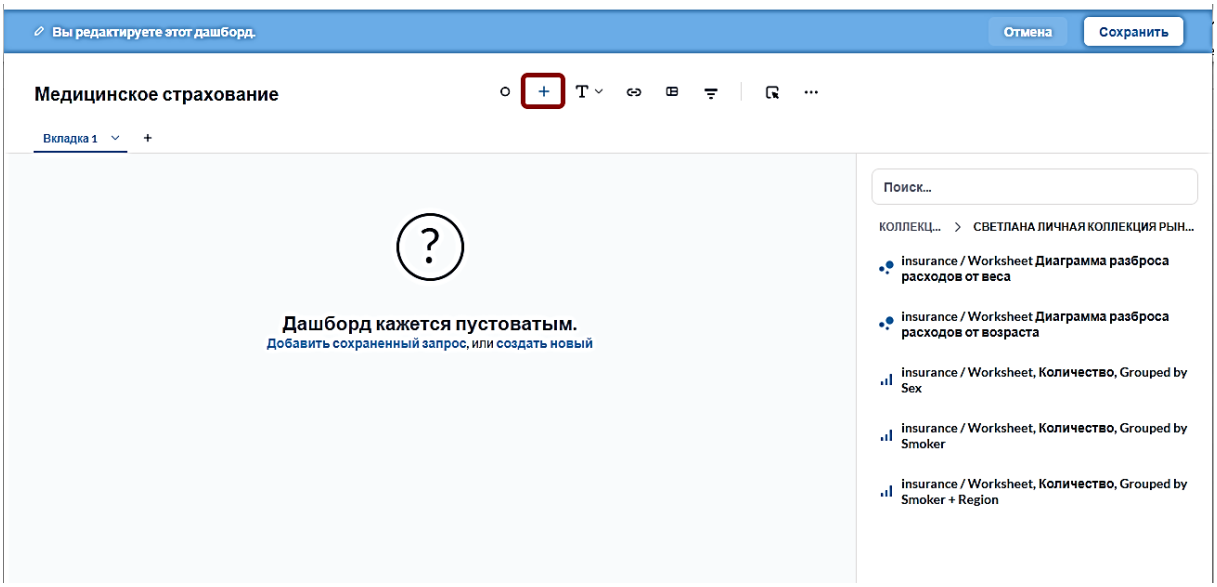


Рис. 1.32. Панель с перечнем созданных запросов

Также на дашборд можно добавить текст или заголовок (рис. 1.33), чтобы сделать его более информативным.

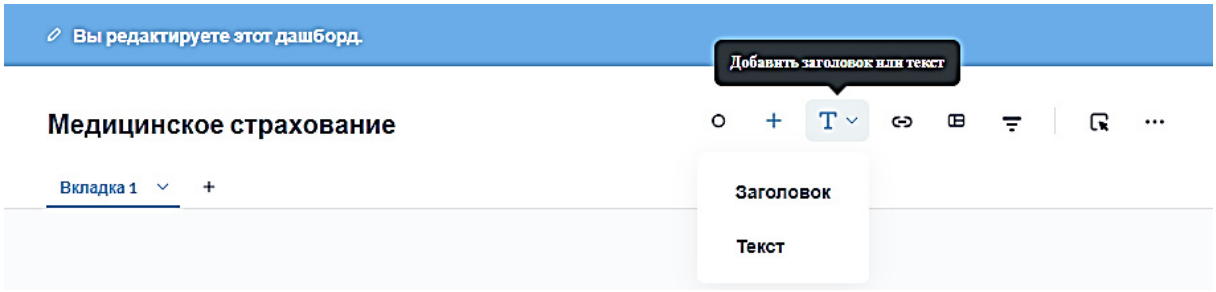


Рис. 1.33. Добавление описания и заголовков на дашборд

Разместим построенные визуализации на новом дашборде (рис. 1.34).

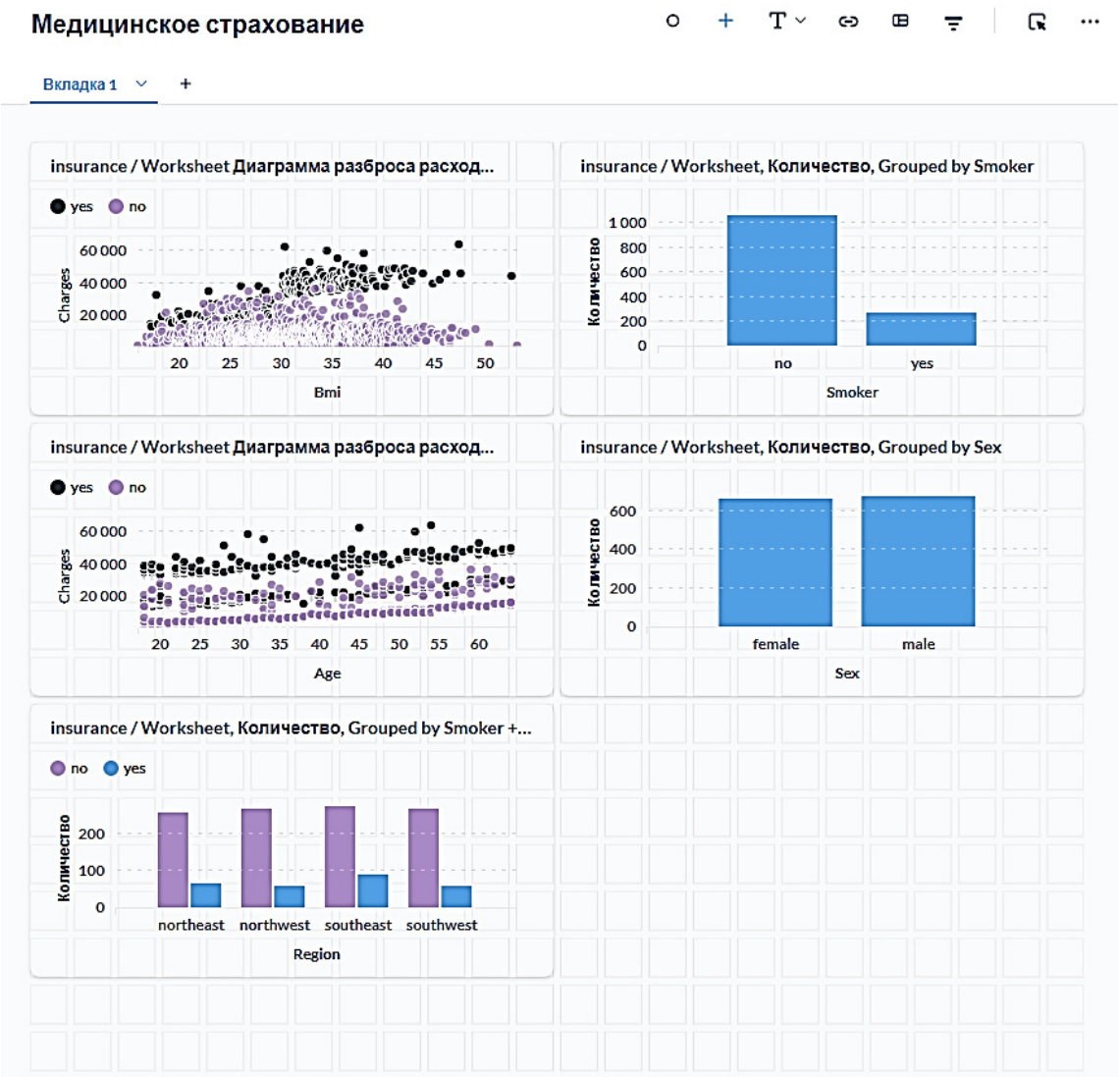


Рис. 1.34. Дашборд, включающий построенные визуализации

Добавим фильтр (рис. 1.35). Сначала выберем для настройки фильтра числовой показатель age (рис. 1.36–1.38). Тип фильтра выбирается между, чтобы иметь возможность рассматривать различные варианты диапазонов значений показателя (рис. 1.38).

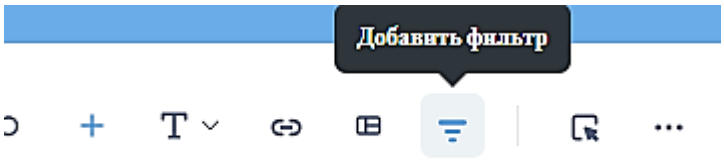


Рис. 1.35. Добавление фильтра

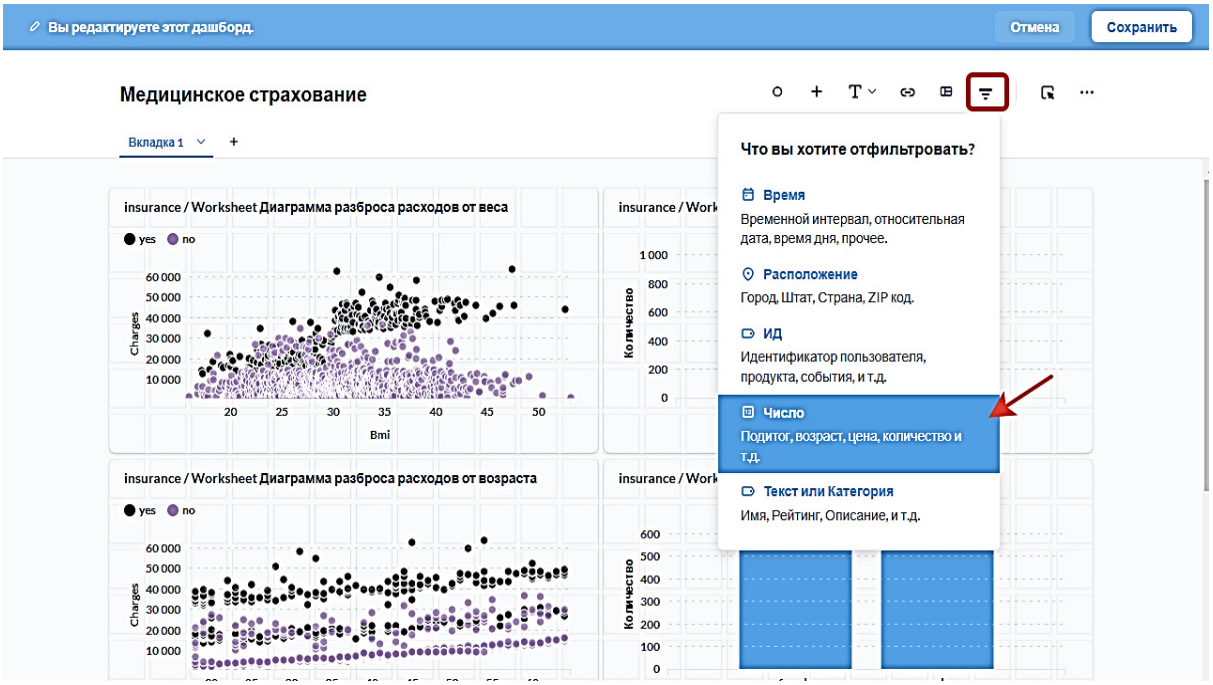


Рис. 1.36. Выбор типа показателя (поля) для настройки фильтра

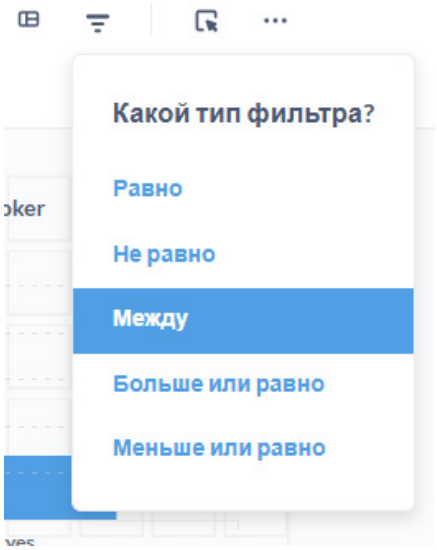


Рис. 1.37. Выбор типа фильтра

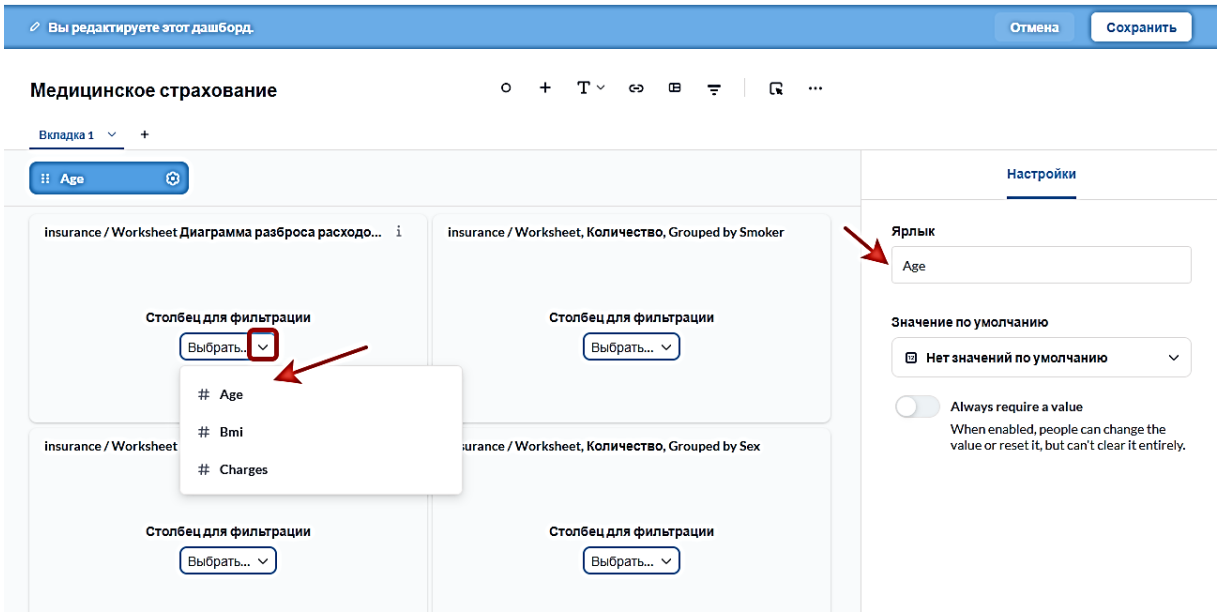


Рис. 1.38. Выбор показателя для фильтрации

В режиме редактирования всегда можно вернуться к настройкам фильтра (рис. 1.39).

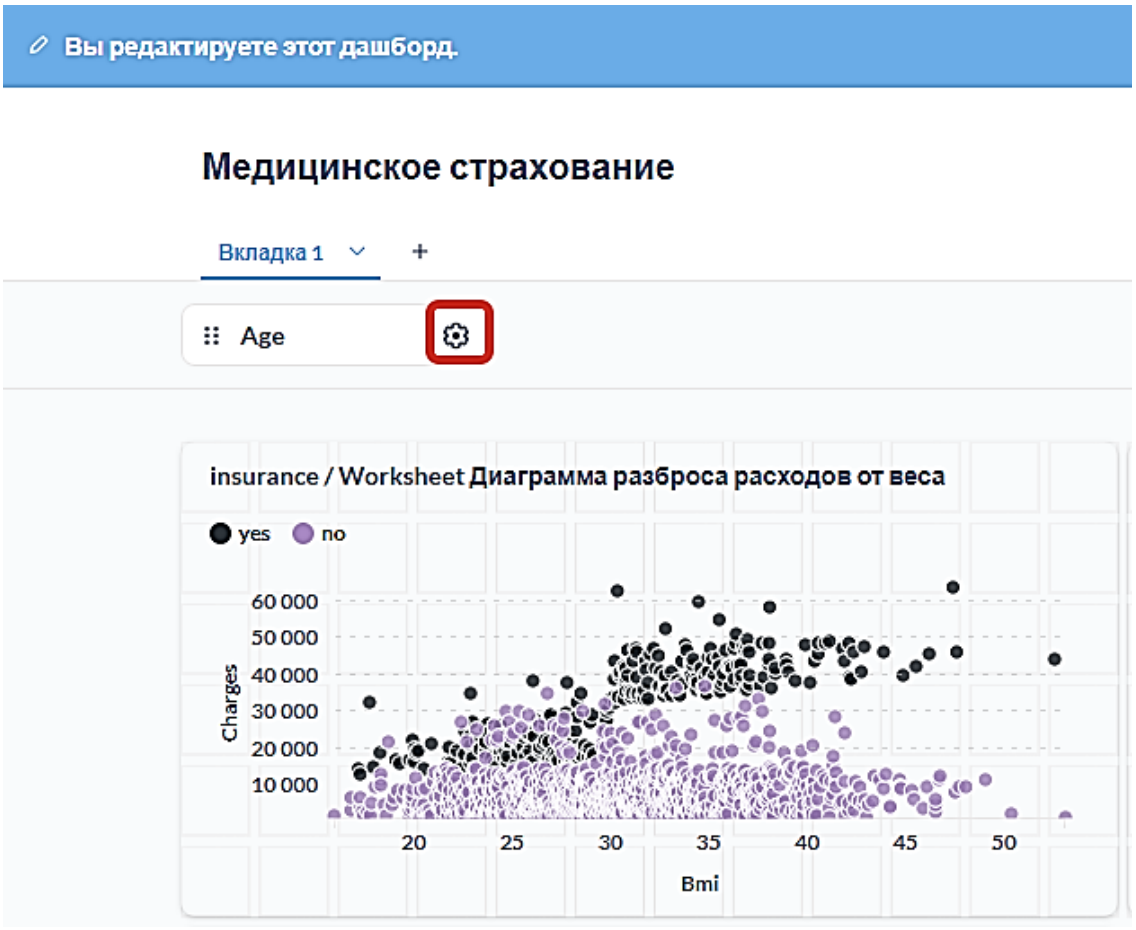


Рис. 1.39. Изменение настроек фильтра (при необходимости)

Также добавим фильтр по показателю region (рис. 1.40–1.42).

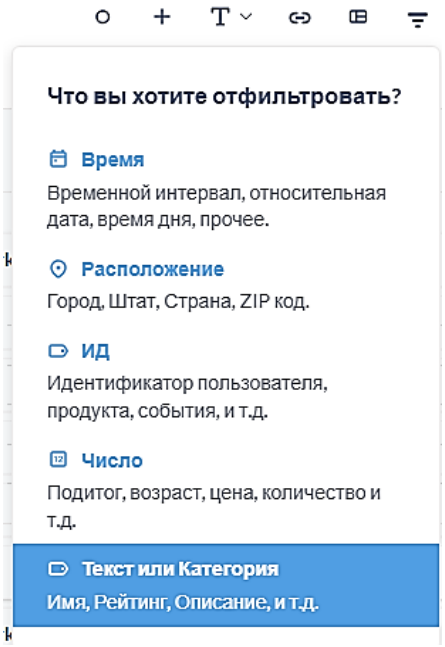


Рис. 1.40. Выбор типа показателя (поля) для настройки фильтра

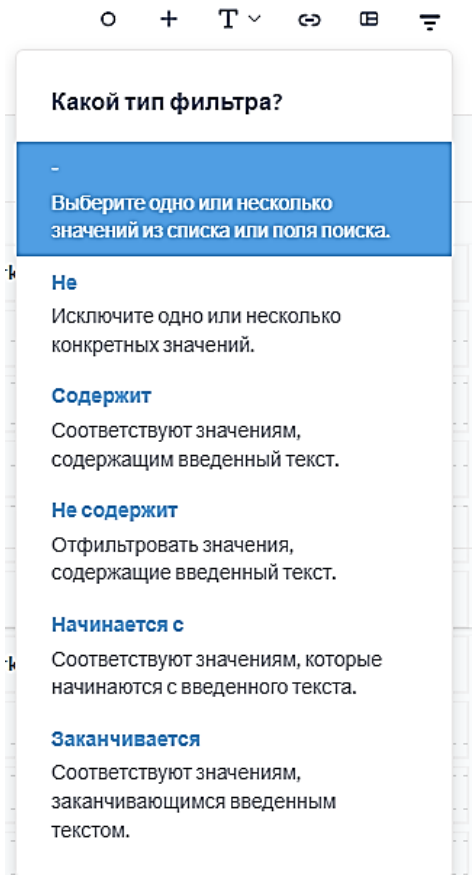


Рис. 1.41. Выбор типа фильтра

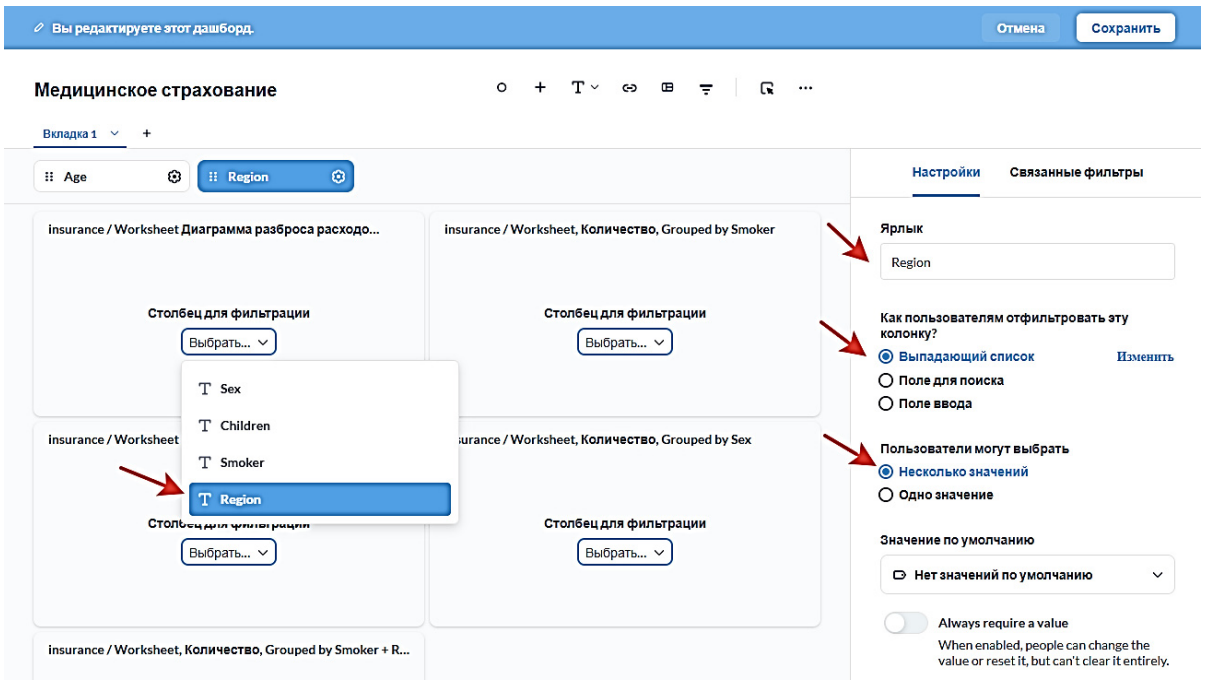


Рис. 1.42. Выбор показателя для фильтрации

При сохранении дашборда автоматически перейдем в режим просмотра (рис. 1.43).

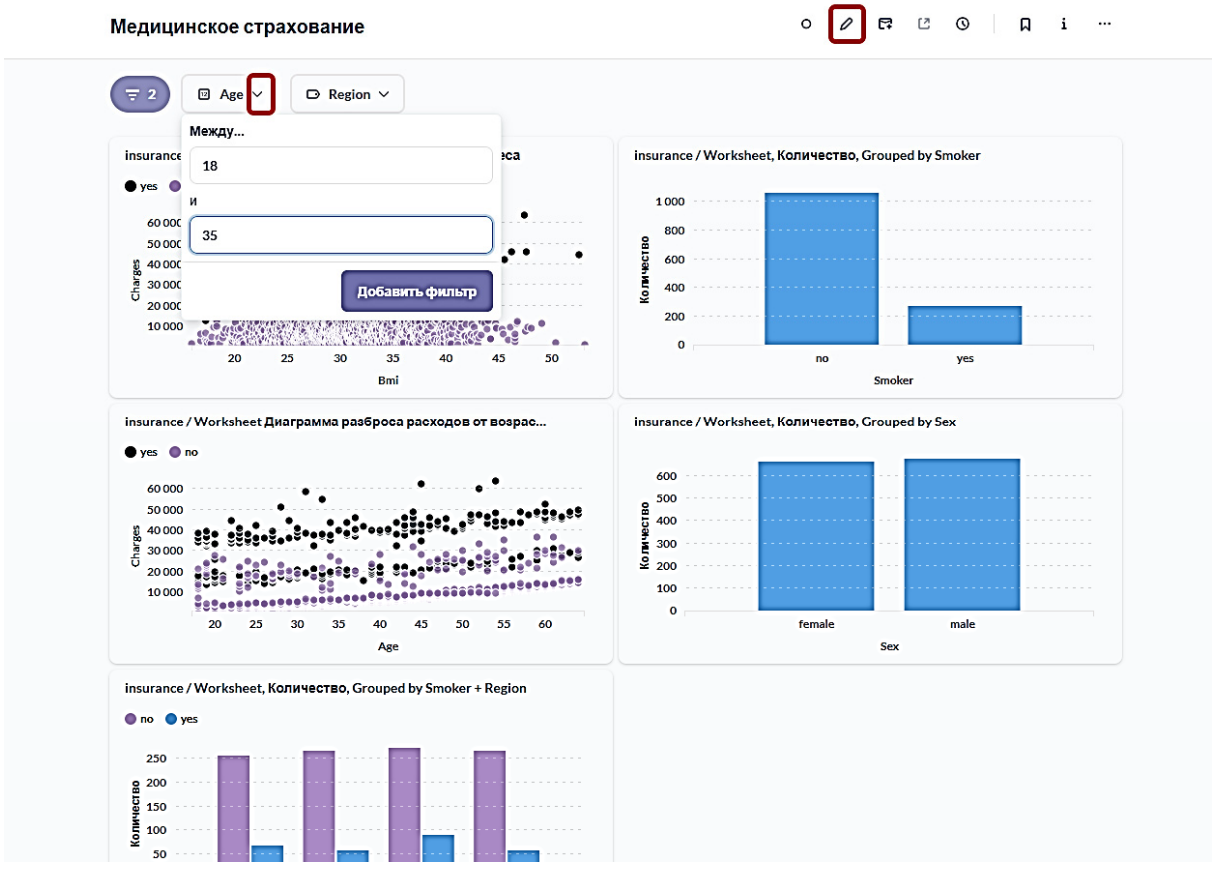


Рис. 1.43. Добавление значений в фильтр по показателю age

Обратно в режим редактирования возвращает кнопка пиктографического меню . Внесем значения в фильтр по показателю age (от 18 до 35) и по показателю region, выбрав три района из четырех (рис. 1.44).

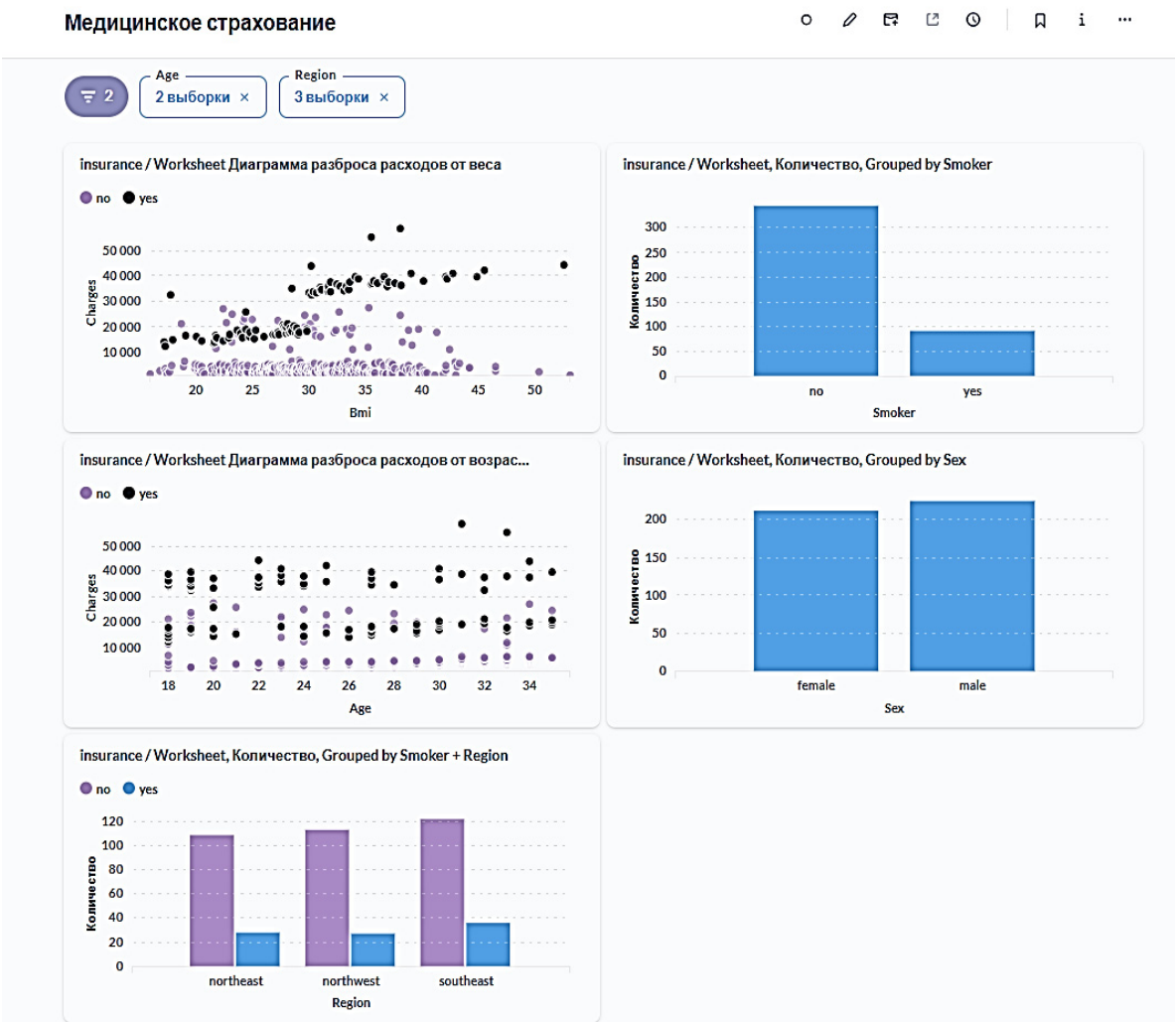
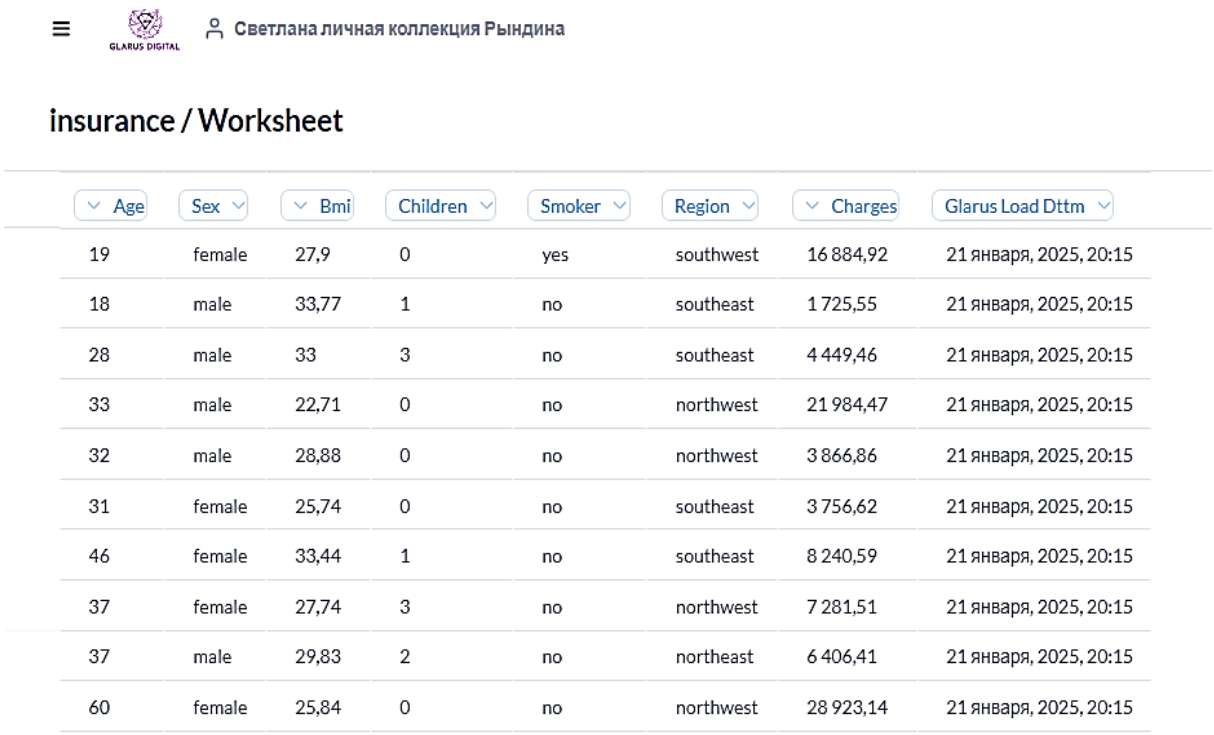


Рис. 1.44. Дашборд с двумя активными фильтрами

Создадим новый запрос по всем данным с визуализацией таблица (рис. 1.45).

На созданный дашборд добавим ссылку (рис. 1.46) на источник данных на ресурсе (https://www.kaggle.com/mirichoi0218/insurance) и на запрос с таблицей данных (рис. 1.47).

В качестве заголовков для визуализаций на дашборде используются названия запросов. Однако в каждой активной визуализации на дашборде в режиме редактирования имеется кнопка для изменения вида визуализации (рис. 1.48), позволяющая менять заголовки, подписи осей и т.п.



Age	Sex	Bmi	Children	Smoker	Region	Charges	Glarus Load Dttm
19	female	27,9	0	yes	southwest	16 884,92	21 января, 2025, 20:15
18	male	33,77	1	no	southeast	1 725,55	21 января, 2025, 20:15
28	male	33	3	no	southeast	4 449,46	21 января, 2025, 20:15
33	male	22,71	0	no	northwest	21 984,47	21 января, 2025, 20:15
32	male	28,88	0	no	northwest	3 866,86	21 января, 2025, 20:15
31	female	25,74	0	no	southeast	3 756,62	21 января, 2025, 20:15
46	female	33,44	1	no	southeast	8 240,59	21 января, 2025, 20:15
37	female	27,74	3	no	northwest	7 281,51	21 января, 2025, 20:15
37	male	29,83	2	no	northeast	6 406,41	21 января, 2025, 20:15
60	female	25,84	0	no	northwest	28 923,14	21 января, 2025, 20:15

Рис. 1.45. Исходные данные набора

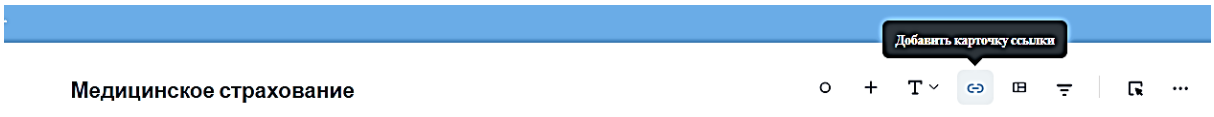


Рис. 1.46. Добавление ссылки

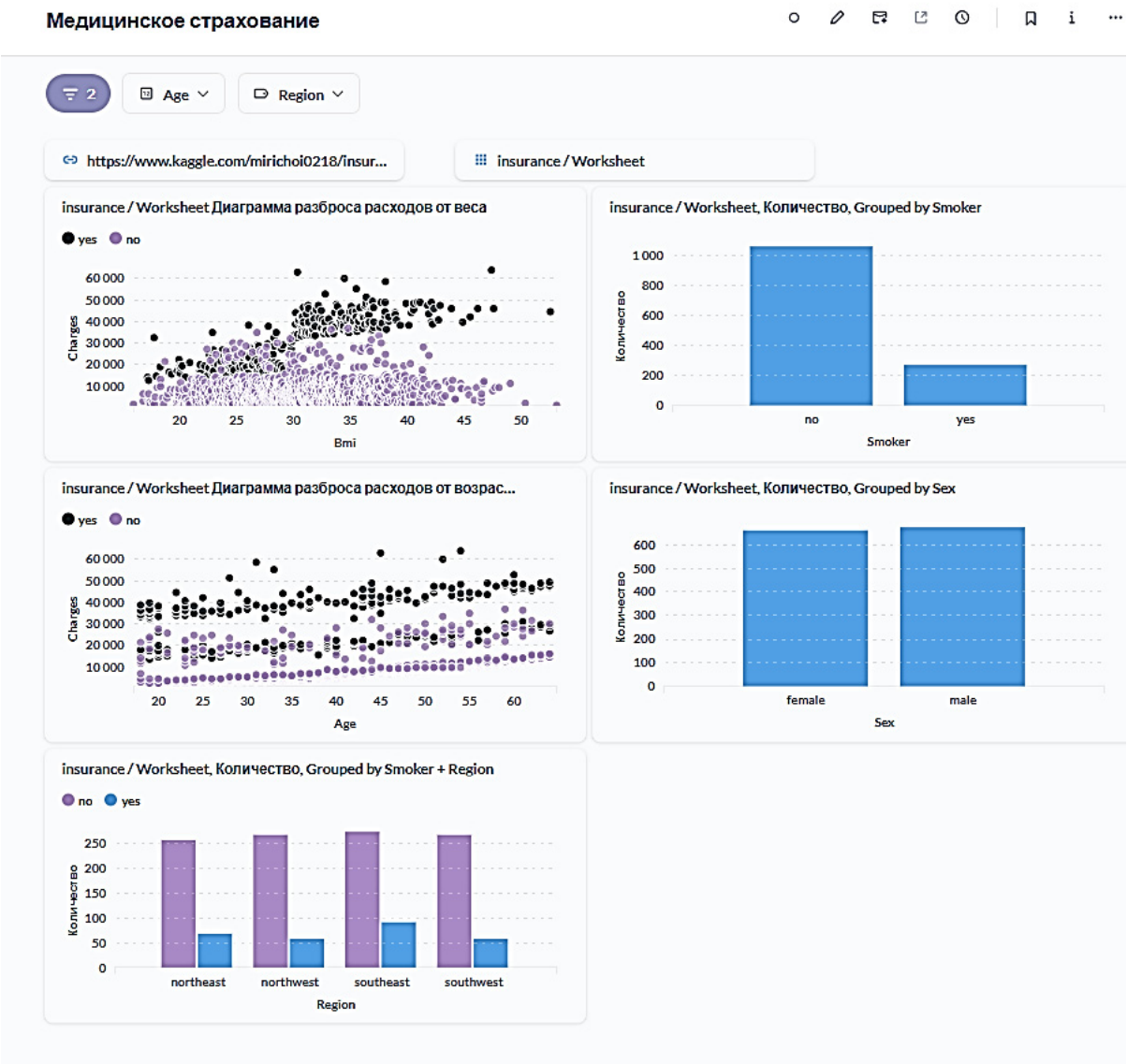


Рис. 1.47. Дашборд со ссылками на данные

Изменим заголовок и цвет фона для одной из визуализаций (рис. 1.49).

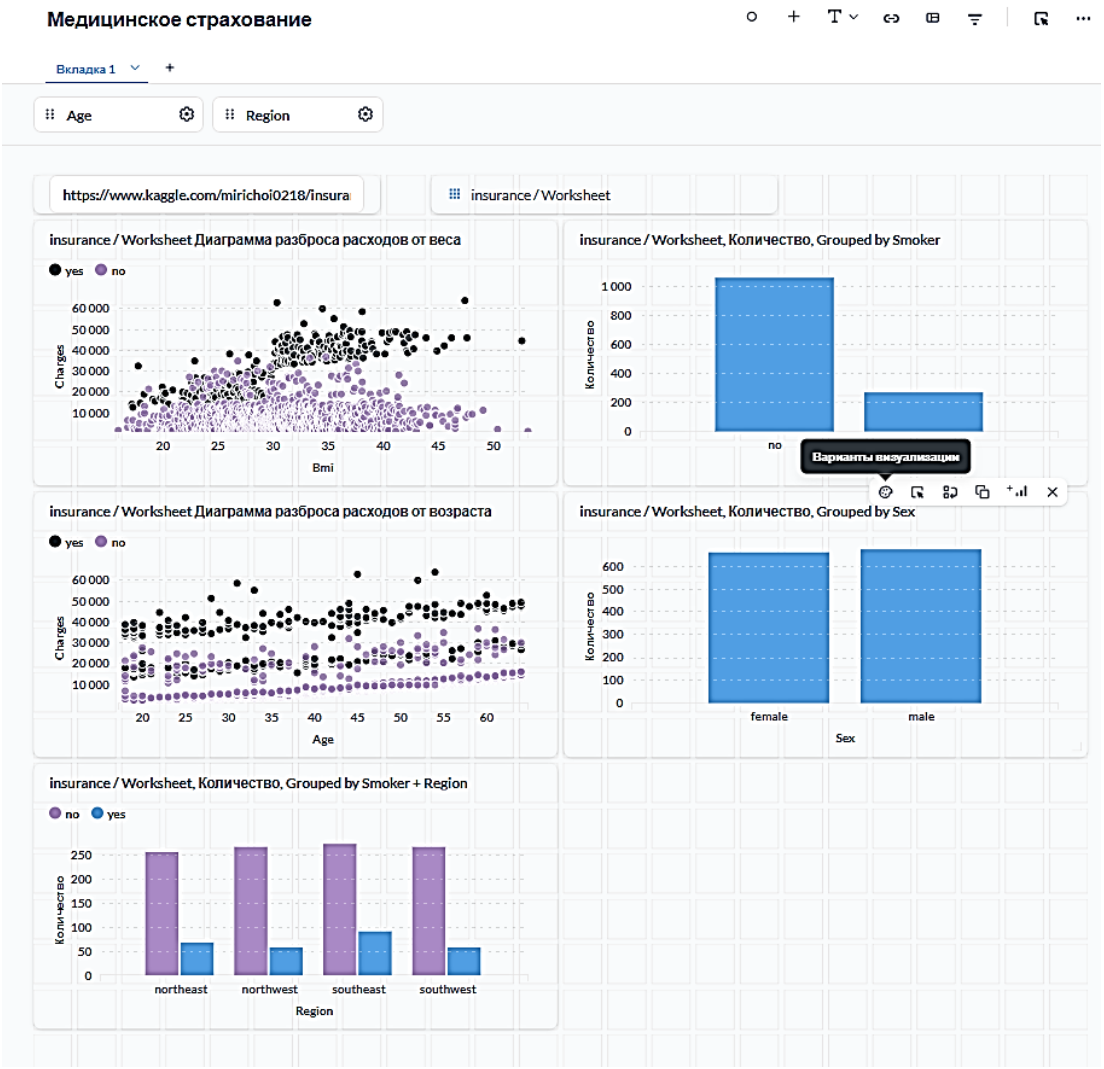


Рис. 1.48. Переход к форматированию визуализации

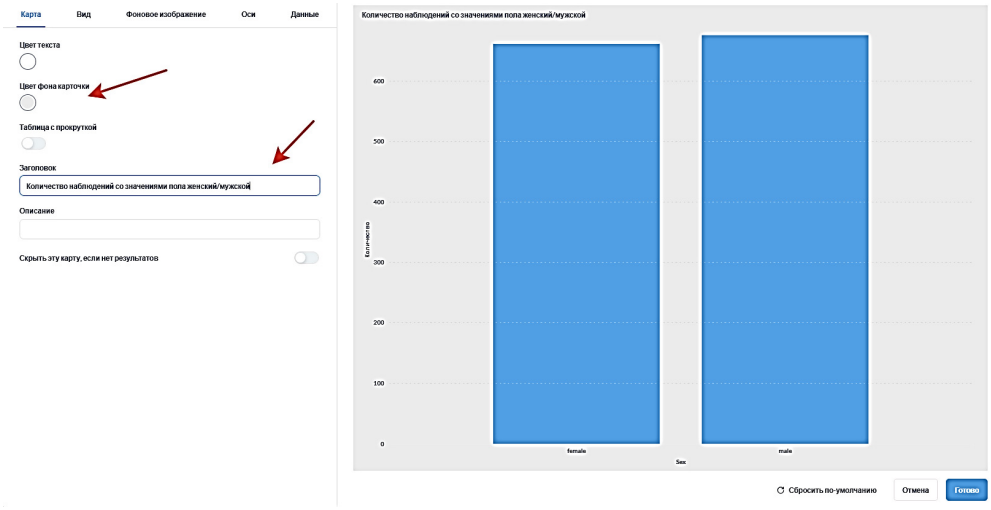


Рис. 1.49. Форматирование визуализации

Аналогично можно отформатировать и остальные визуализации.

1.5. Объединение данных и визуализаций

Создадим новую коллекцию с именем Продажи (рис. 1.50), в которой будем размещать запросы по новым наборам данных.

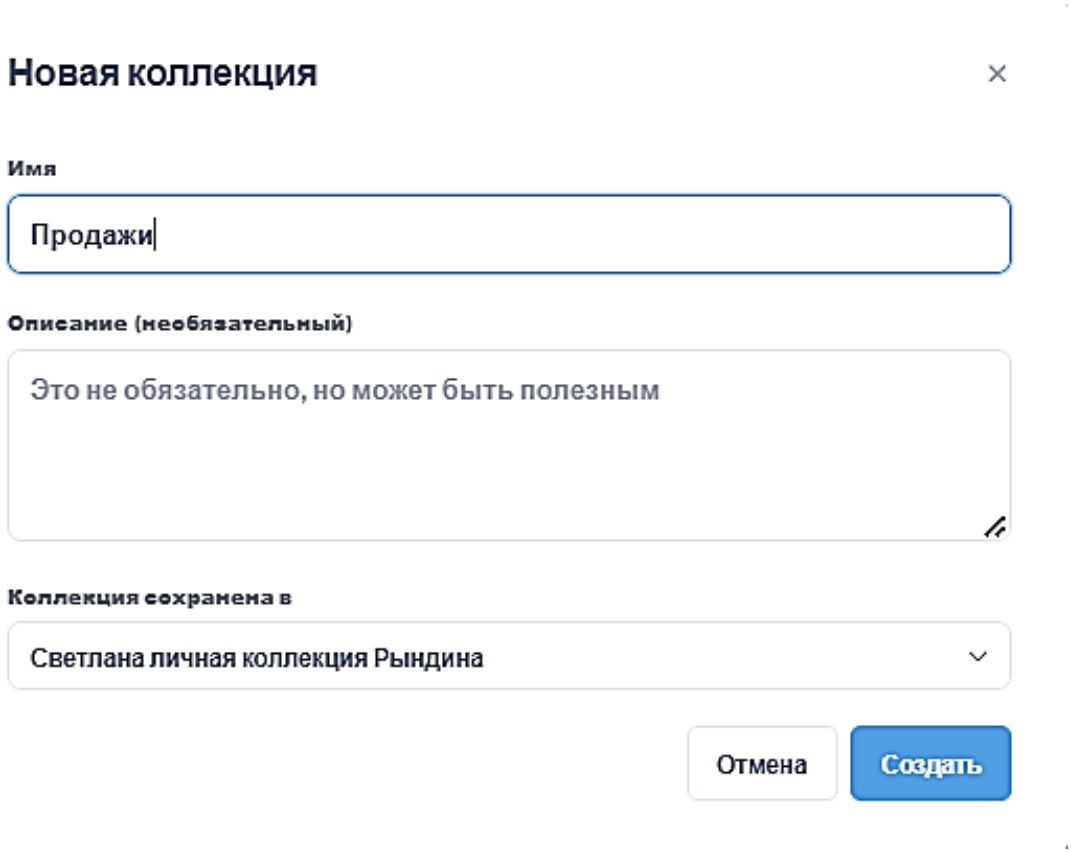


Рис. 1.50. Создание коллекции

Наборы данных, которые будут использоваться: данные о транзакциях (https://www.kaggle.com/datasets/regivm/retailtransactiondata?select=Retail_Data_Transactions.csv) и об отклике тех же клиентов на рекламные взаимодействия (https://www.kaggle.com/datasets/regivm/retailtransactiondata?select=Retail_Data_Response.csv).

Переведем эти наборы в нужный формат файла с помощью MS Excel. Сначала импортируем в созданный файл в MS Excel данные из Retail_Data_Transactions.csv (рис. 1.51), изменив представление для даты, если потребуется (рис. 1.52).

Аналогично импортируем в тот же файл данные из Retail_Data_Response.csv и удалим все остальные листы (рис. 1.53).

Сохраним файл с именем retail.xlsx. И загрузим этот файл на платформу. Он будет загружен в виде двух таблиц (рис. 1.54).

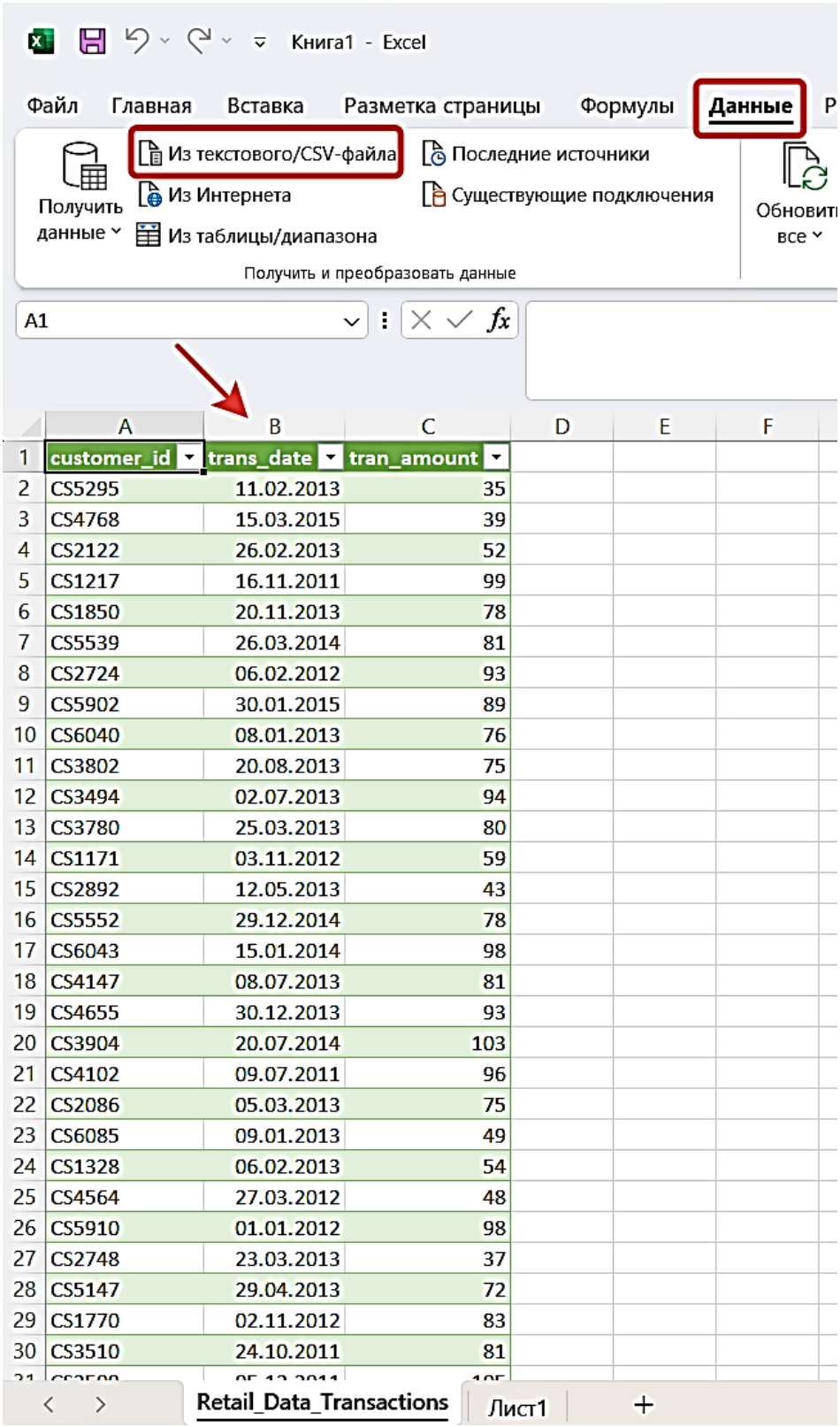


Рис. 1.51. Импорт данных из текстового файла

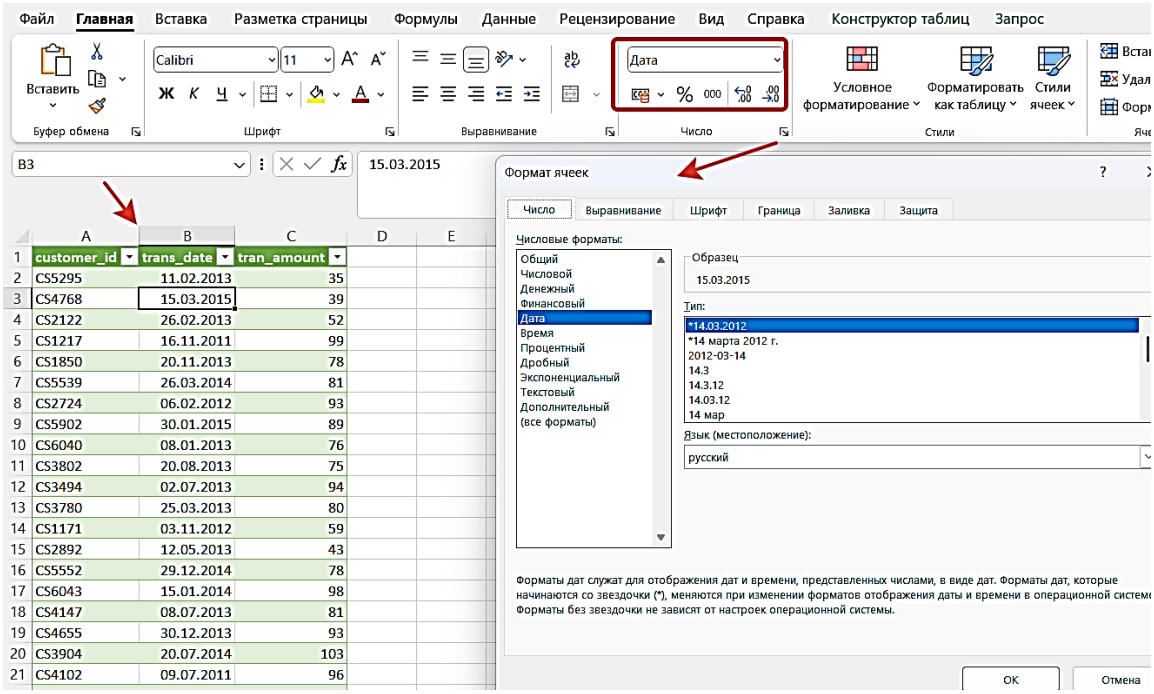


Рис. 1.52. Форматирование даты

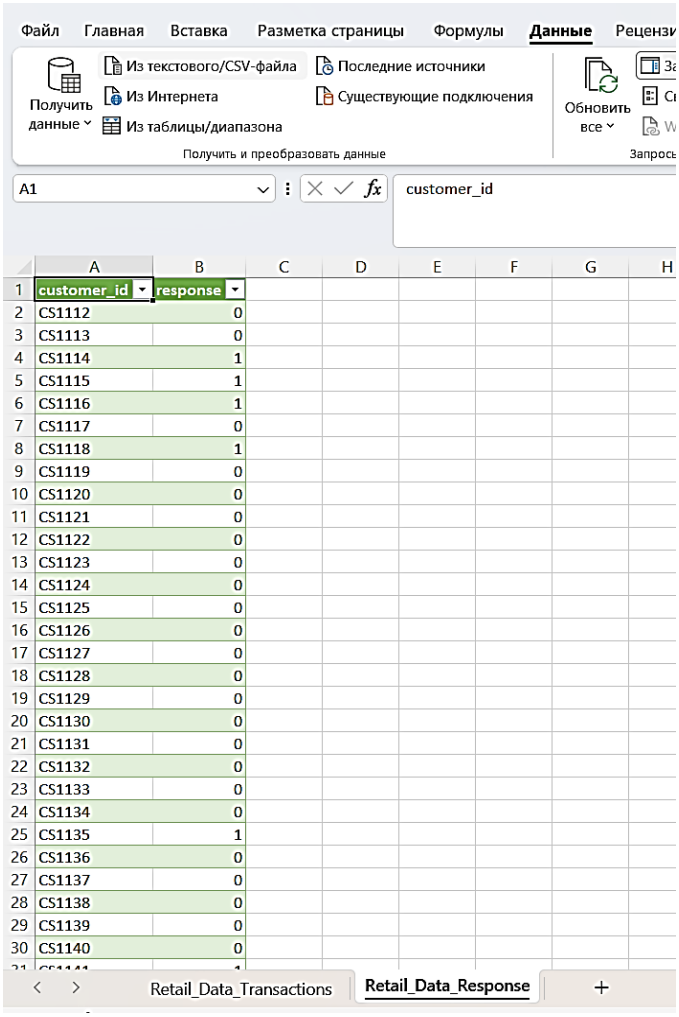


Рис. 1.53. Импорт данных из второго файла

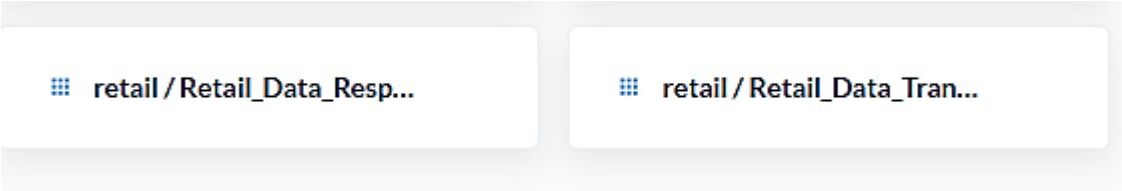


Рис. 1.54. Результат загрузки файла retail.xlsx на платформу

Откроем сведения о первой таблице Retail_Data_Transactions (рис. 1.55).

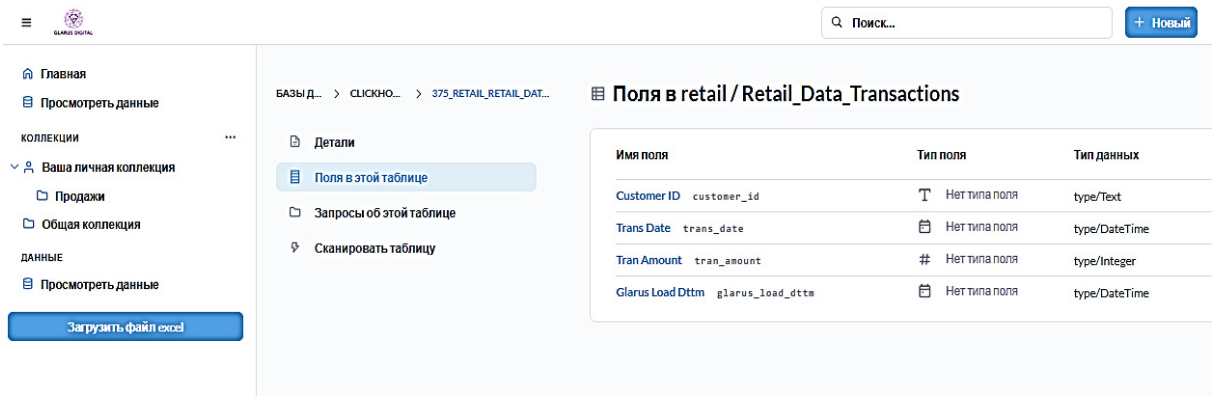


Рис. 1.55. Сведения о наборе данных Retail_Data_Transactions

Ранее в данных о медицинском страховании показатель children определялся как категориальный с типом integer, в этих данных tran_amount определяется как числовой с типом integer. Тип поля в этом случае зависит от числа уникальных значений по сравнению с числом наблюдений, если уникальных значений мало, то показатель считается категориальным, иначе – числовым.

Таблица Retail_Data_Transactions содержит следующие данные:

– customer_id – идентификатор клиента (текстовая/символьная переменная);

– trans_date – дата транзакции;

– tran_amount – сумма покупки.

Таблица Retail_Data_Response содержит следующие данные:

– customer_id – идентификатор клиента (текстовая/символьная переменная);

– response – наличие/отсутствие отклика на рекламную компанию (бинарная переменная отклика).

Создадим запрос, который сравнивает общую сумму продаж клиентам, откликнувшимся на рекламу, с общей суммой продаж тем, кто рекламу проигнорировал.

Для этого к данным первой таблицы Retail_Data_Transactions присоединим данные второй таблицы Retail_Data_Response для изменения типа соединения, кликнув по значку соединения и выбрав внутреннее (рис. 1.56). Также выберем поля этих таблиц, по которым объединение произойдет (Customer_ID). Важно, чтобы поля, по которым происходит соединение таблиц, имели один и тот же тип данных и содержали одинаковые значения. Если рядом с таблицей запроса появляется стрелка, раскрывающая список полей, то в ней можно настроить поля для отображения (рис. 1.57).

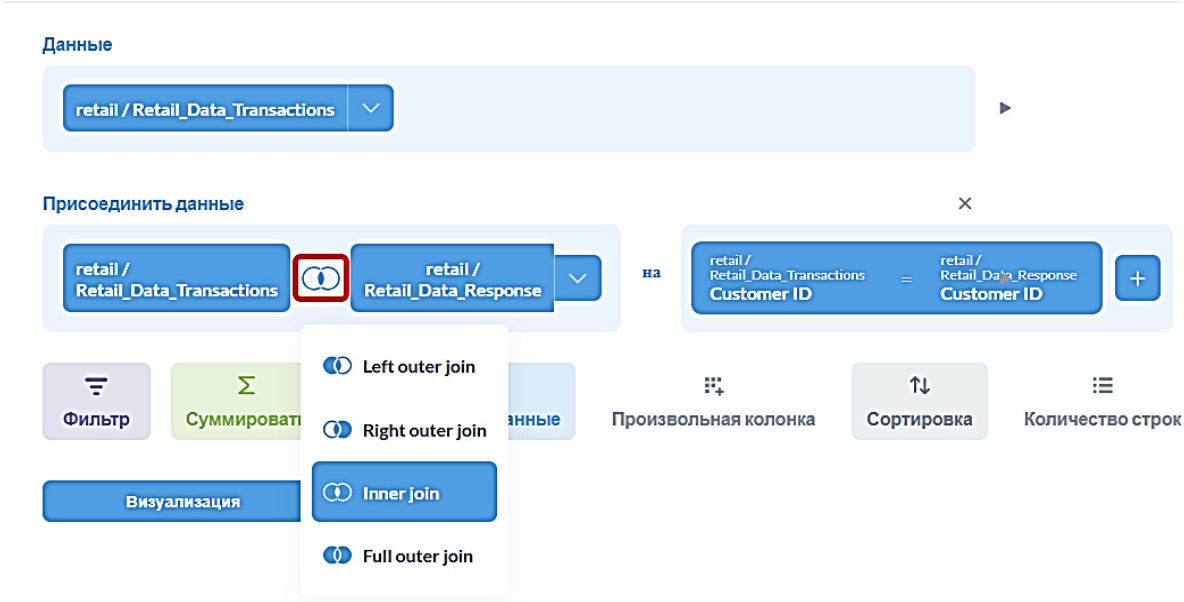


Рис. 1.56. Изменение типа соединения

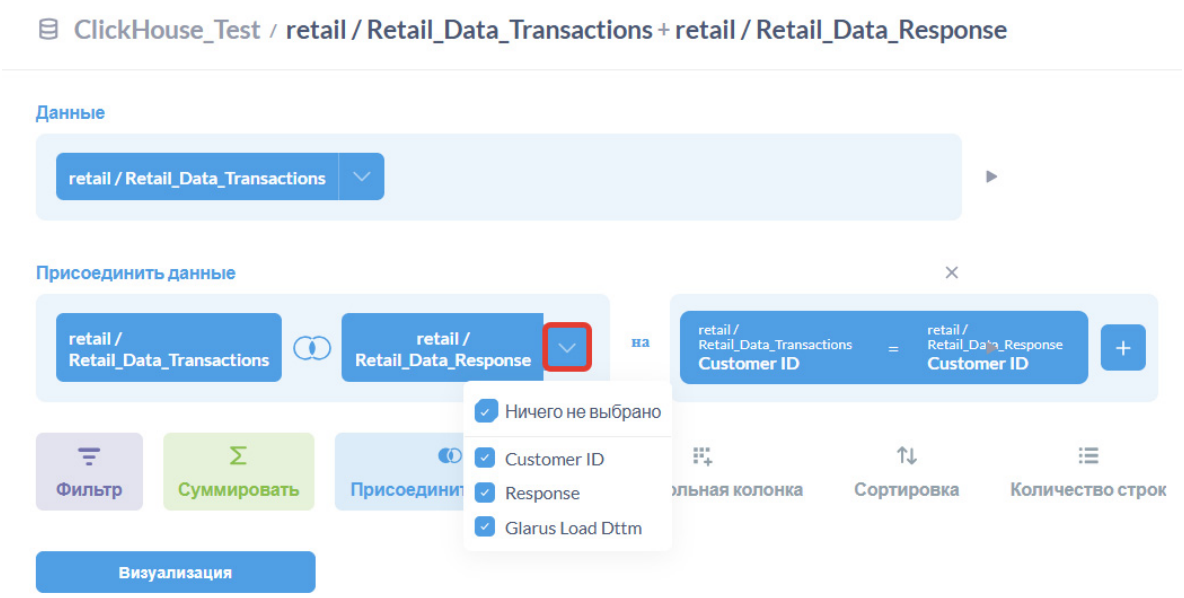


Рис. 1.57. Выбор полей

Агрегацию проведем по полю tran_amount первой таблицы Retail_Data_Transactions, а группировать будем по переменной response второй таблицы Retail_Data_Response (рис. 1.58).

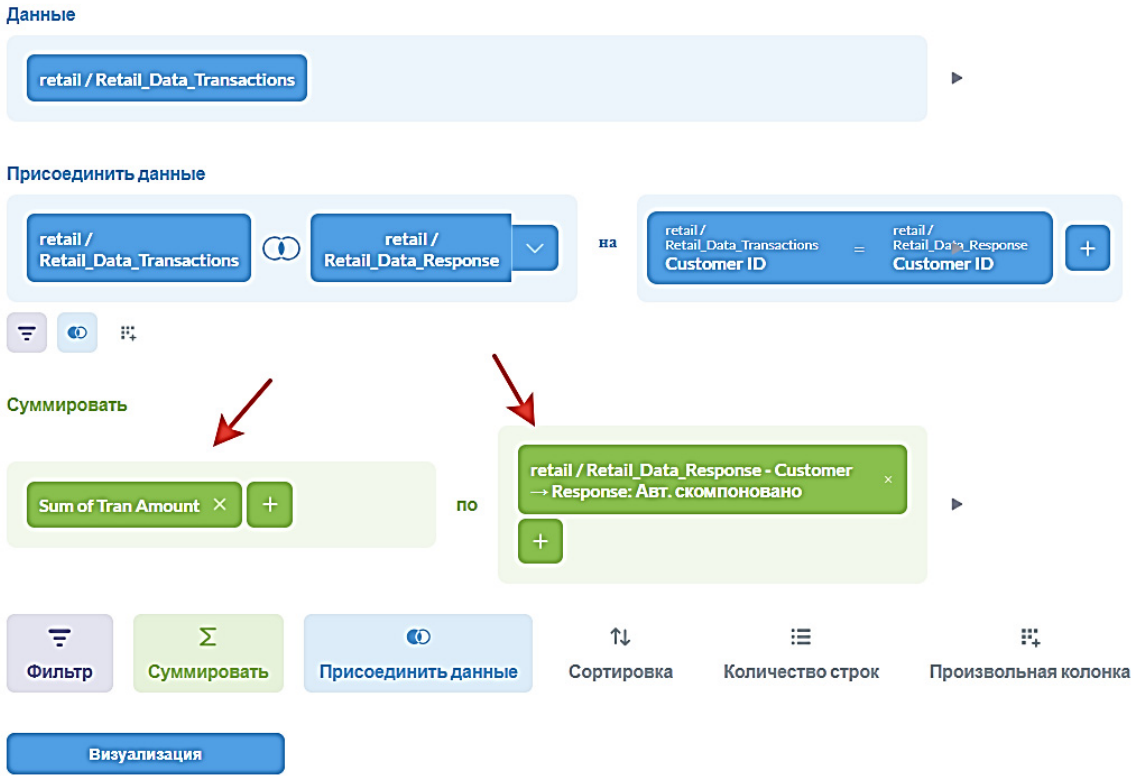


Рис. 1.58. Настройки для запроса по данным двух объединенных таблиц

Построим визуализацию (гистограмму) и сохраним запрос в созданную коллекцию Продажи, сократив его название (рис. 1.59).

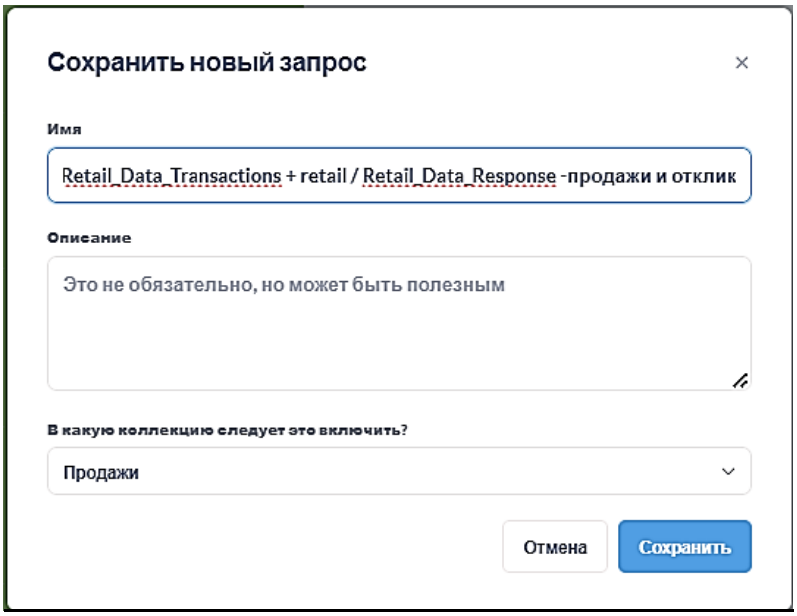


Рис. 1.59. Сохранение запроса

Построим гистограмму по числу клиентов, сгруппированному в зависимости от отклика на рекламу (рис. 1.60, 1.61).

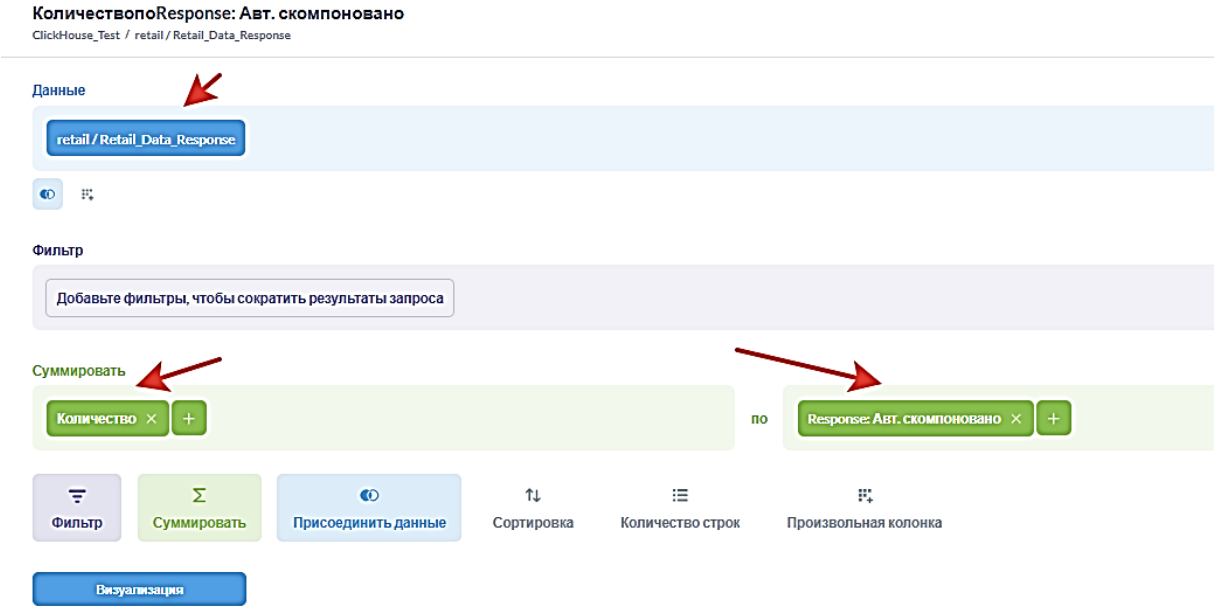


Рис. 1.60. Настройки запроса

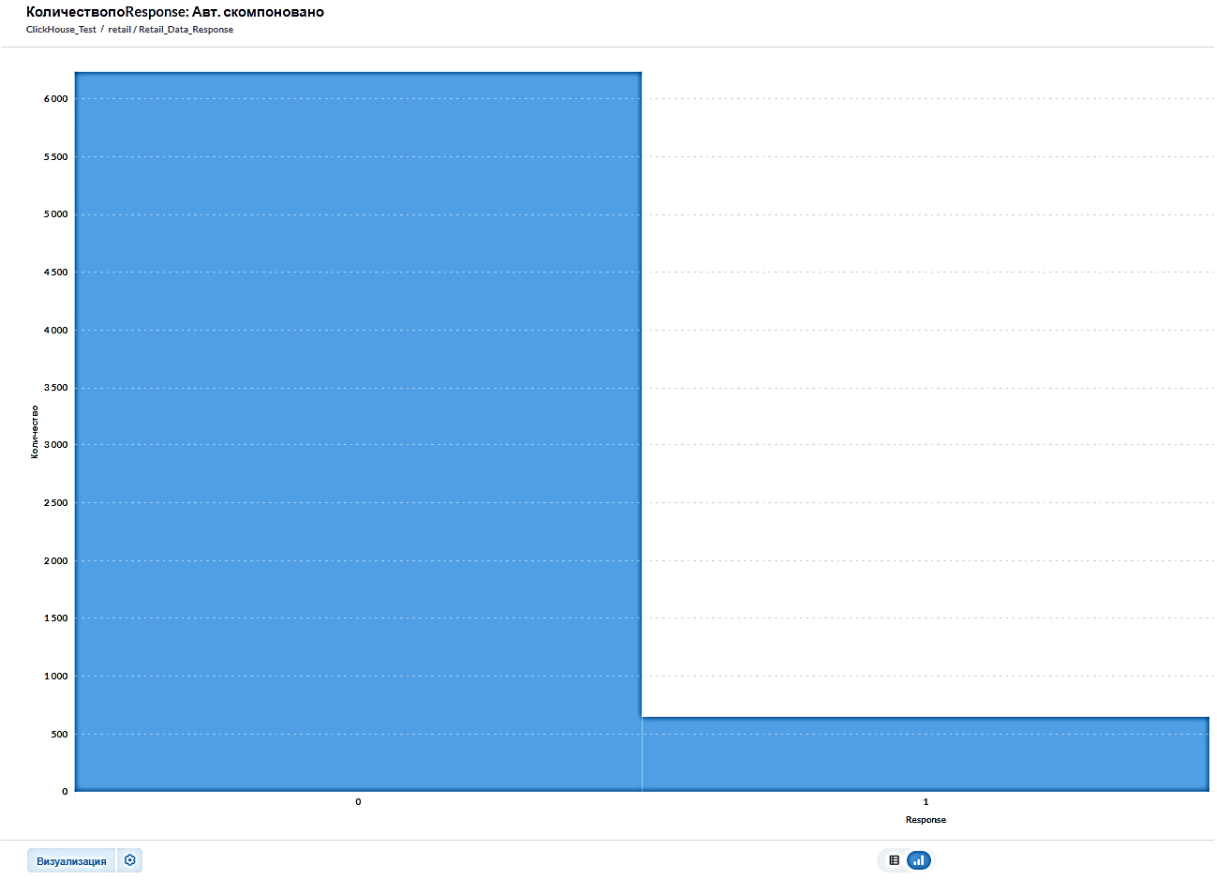


Рис. 1.61. Построенная визуализация

Создадим новый дашборд с именем Продажи (рис. 1.62).

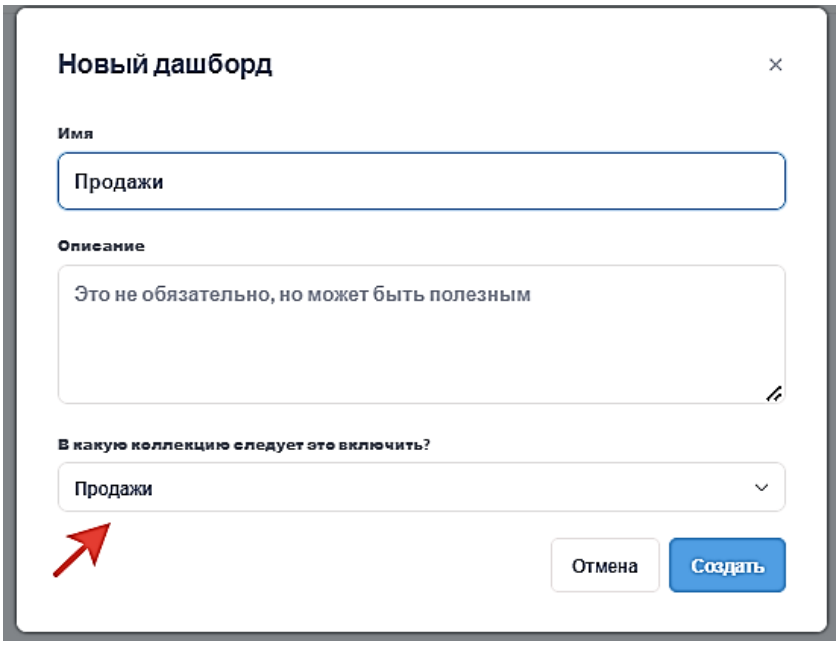


Рис. 1.62. Создание дашборда Продажи

Разместим на дашборде гистограмму, построенную по количеству наблюдений, соответствующих каждому из значений переменной response (рис. 1.63). И в пиктографическом меню этой визуализации выберем значок (Добавить серии), который позволяет дополнить диаграмму другой визуализацией, если в объединяемых визуализациях измерения оси абсцисс совпадают.

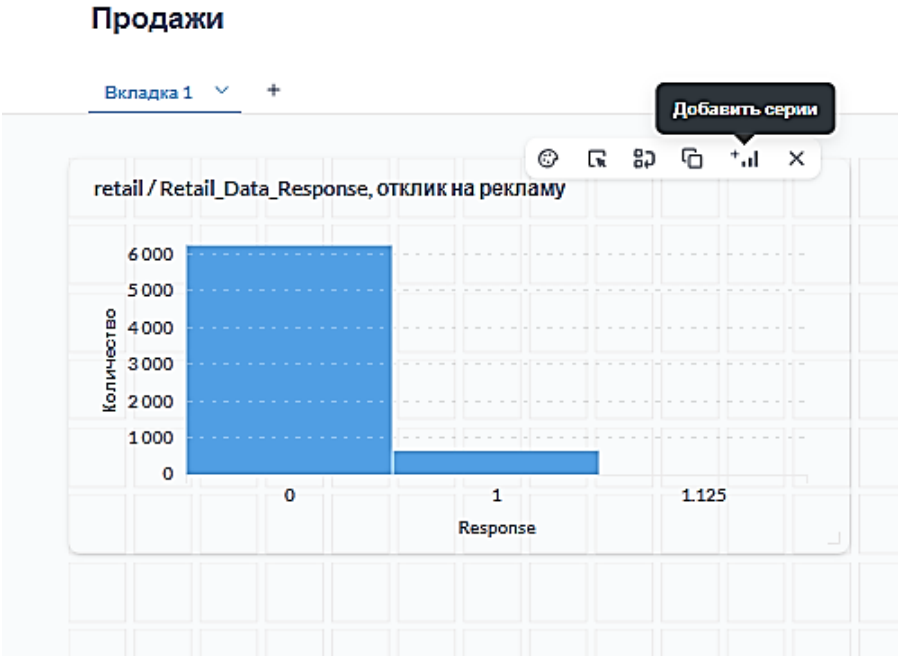


Рис. 1.63. Добавление визуализации на новый дашборд

В диалоговом окне для опции Добавить серии будут отображаться те диаграммы, которые можно объединить с текущей (рис. 1.64).

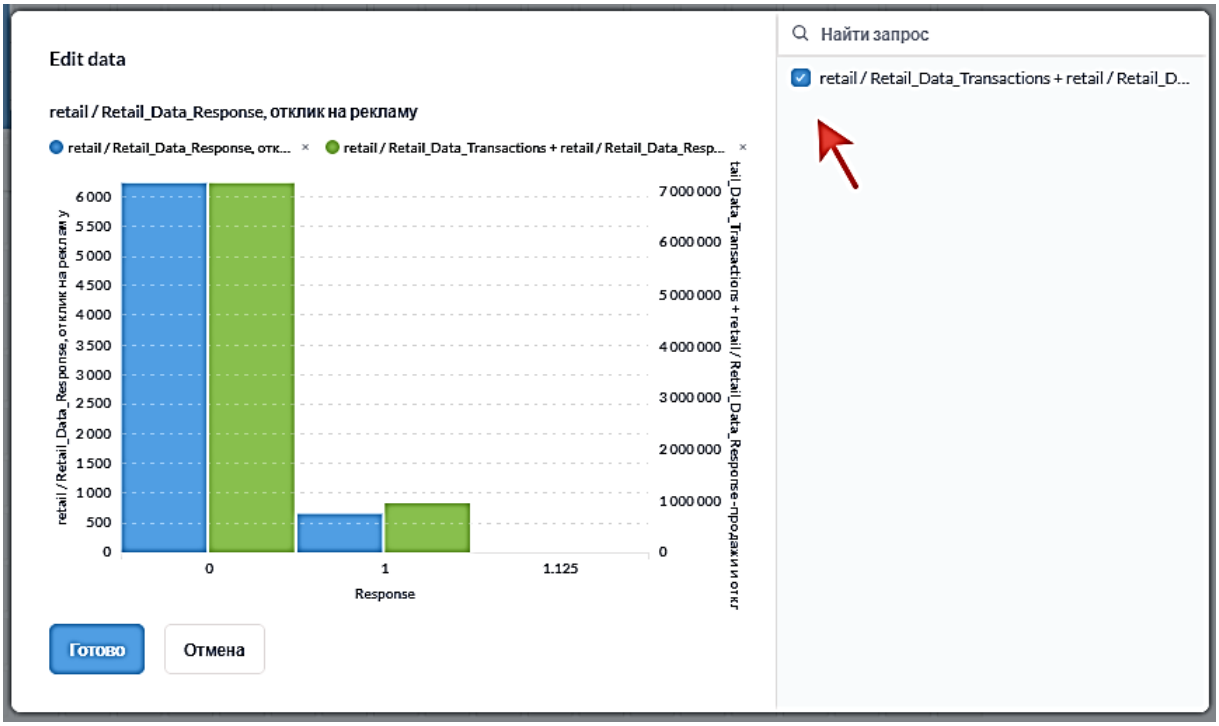


Рис. 1.64. Добавление серии на диаграмму

Скорректируем настройки для объединенных диаграмм (рис. 1.65). Как видим, откликнувшиеся на рекламу покупают больше, просто в общем числе клиентов их доля невелика.

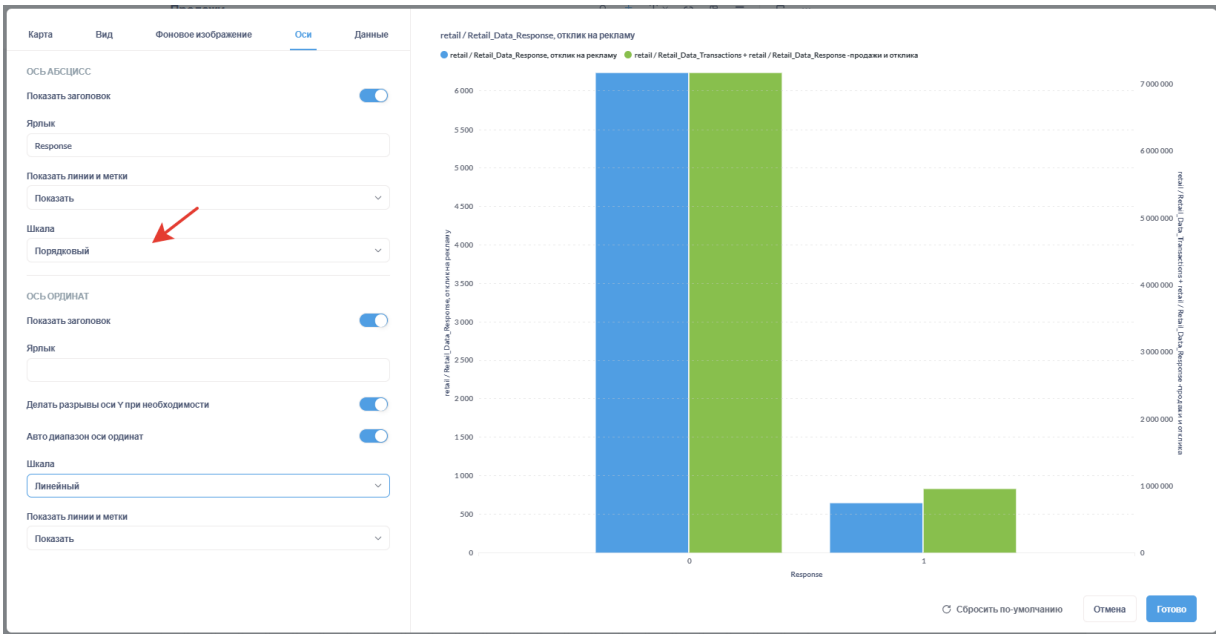


Рис. 1.65. Форматирование объединенной диаграммы (изменение шкалы)

Сохраним дашборд и создадим еще две визуализации: линейную диаграмму по продажам, сгруппированным по месяцам, и таблицу с данными о продажах самым активным клиентам.

Первый из двух новых запросов будет иметь настройки как на рис. 1.66. Важно, что поле с типом данных DateTime, выбранное для группировки, позволяет настроить различные диапазоны для агрегации.

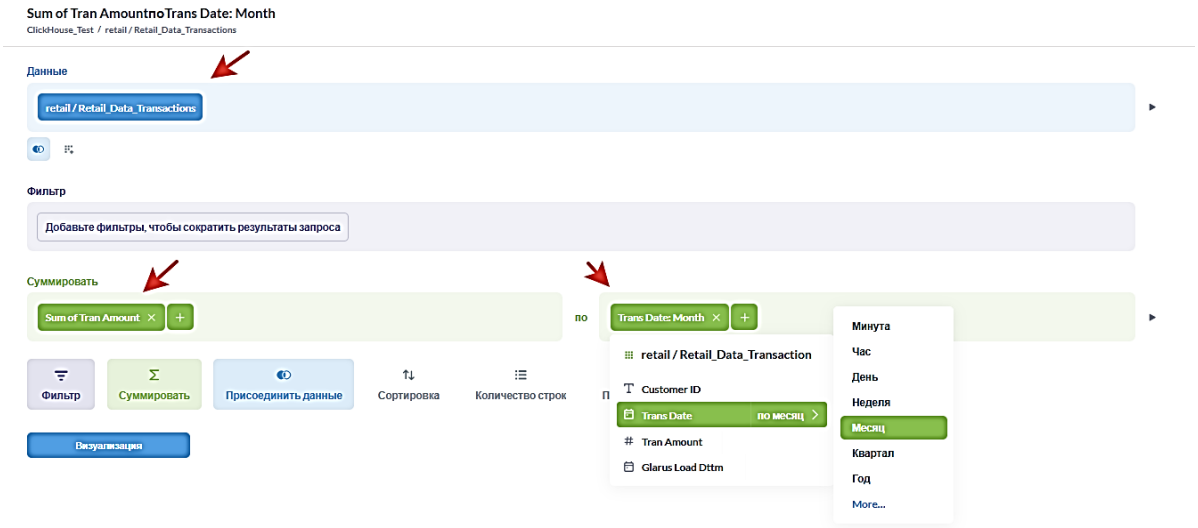


Рис. 1.66. Настройки запроса

Полученная визуализация будет иметь вид, показанный на рис. 1.67.

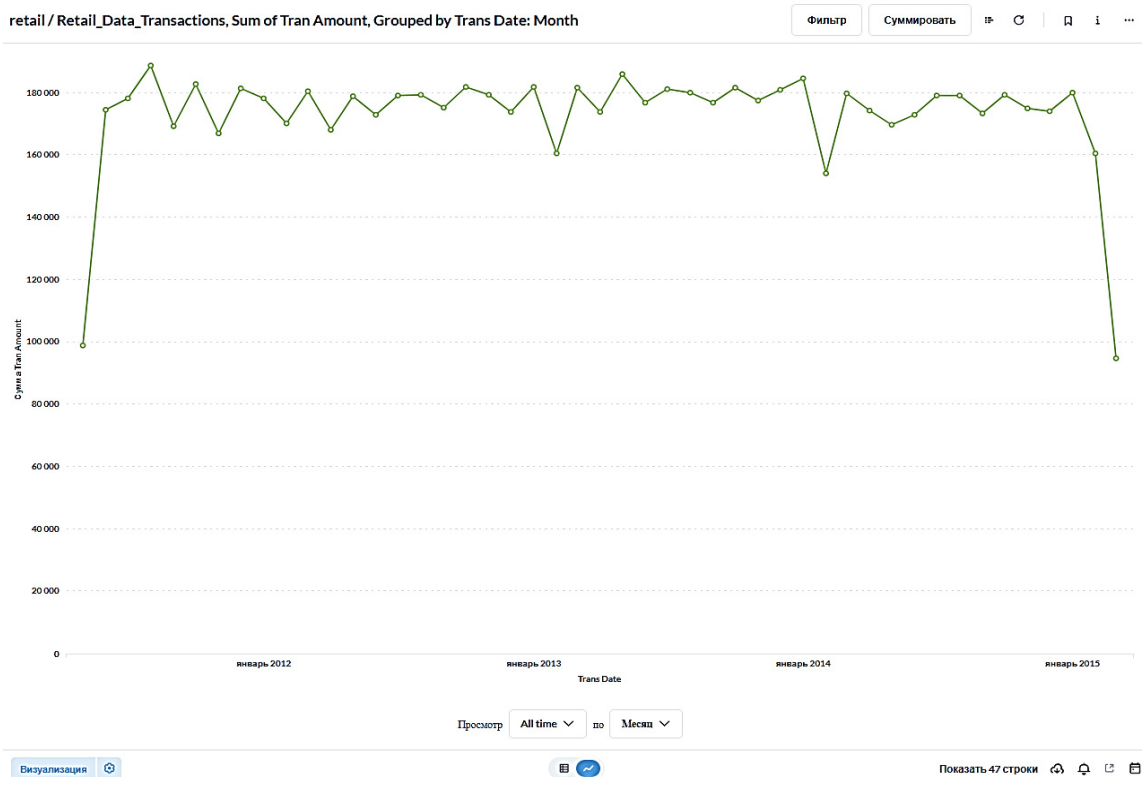


Рис. 1.67. Визуализация по запросу

Для построения второй визуализации просуммируем все покупки по каждому из клиентов и отсортируем полученную таблицу по сумме покупок по убыванию, а также выведем результат запроса на просмотр (рис. 1.68). Это необходимо для настройки фильтра, чтобы установить границу отсечения для суммы покупки. Как видим, сумма больше значения 2400 позволит оставить в рассмотрении только активных клиентов.

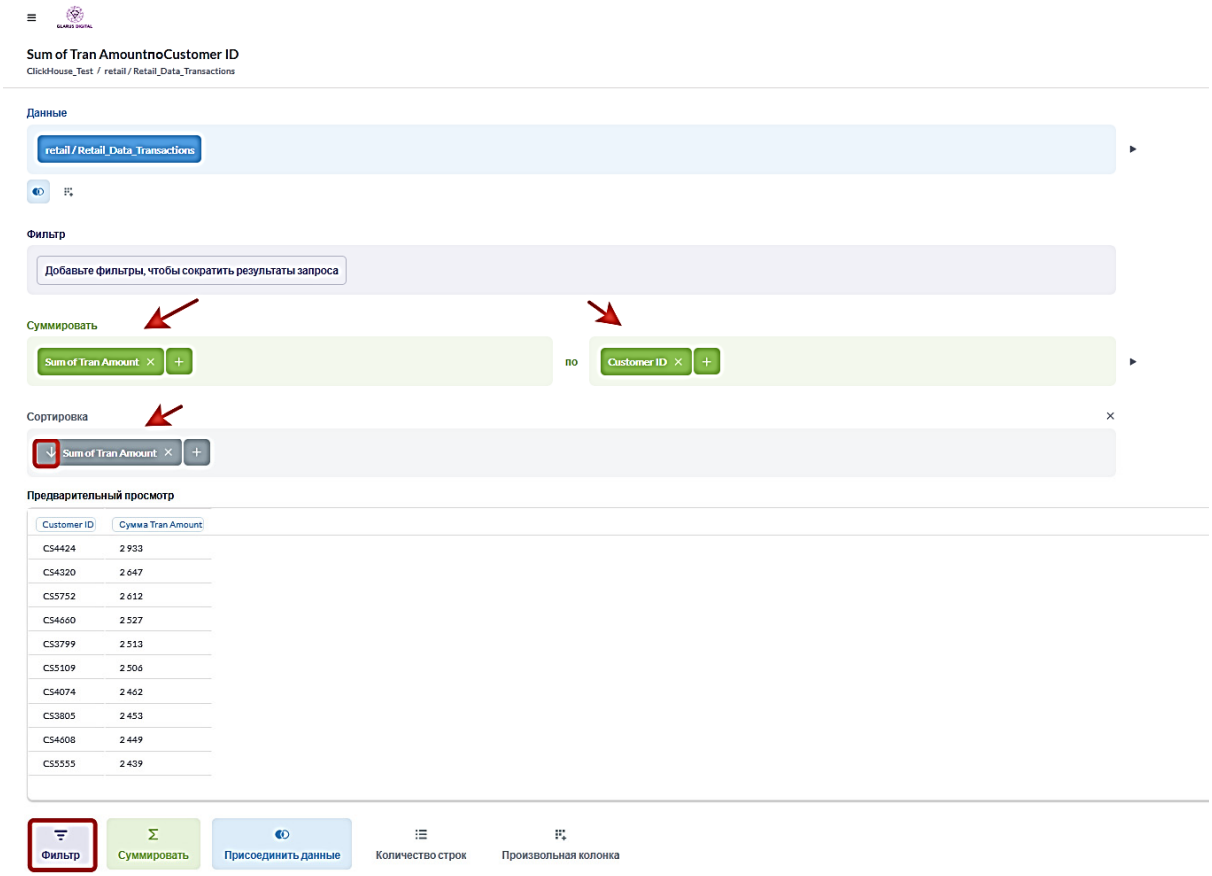


Рис. 1.68. Настройки запроса

Добавим фильтр, выбрав условие Больше чем и значение 2400 (рис. 1.69).

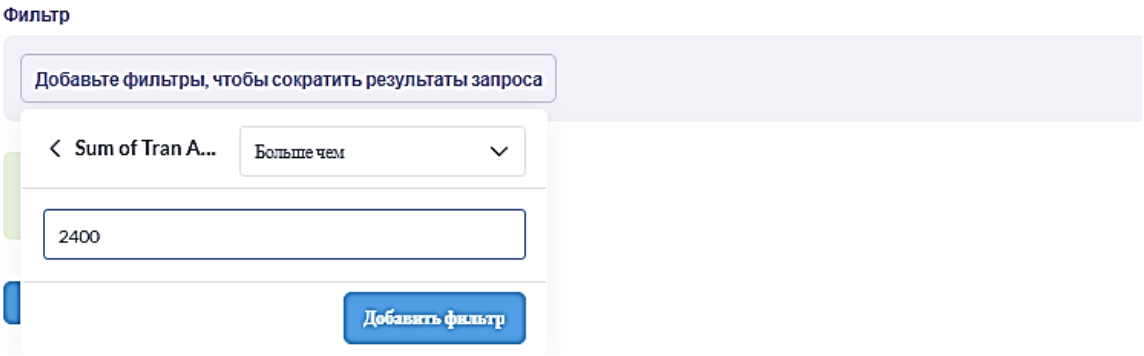


Рис. 1.69. Настройки для фильтра

Добавим две построенные визуализации на дашборд Продажи и скорректируем их заголовки, а также фон одной из визуализаций (рис. 1.70).

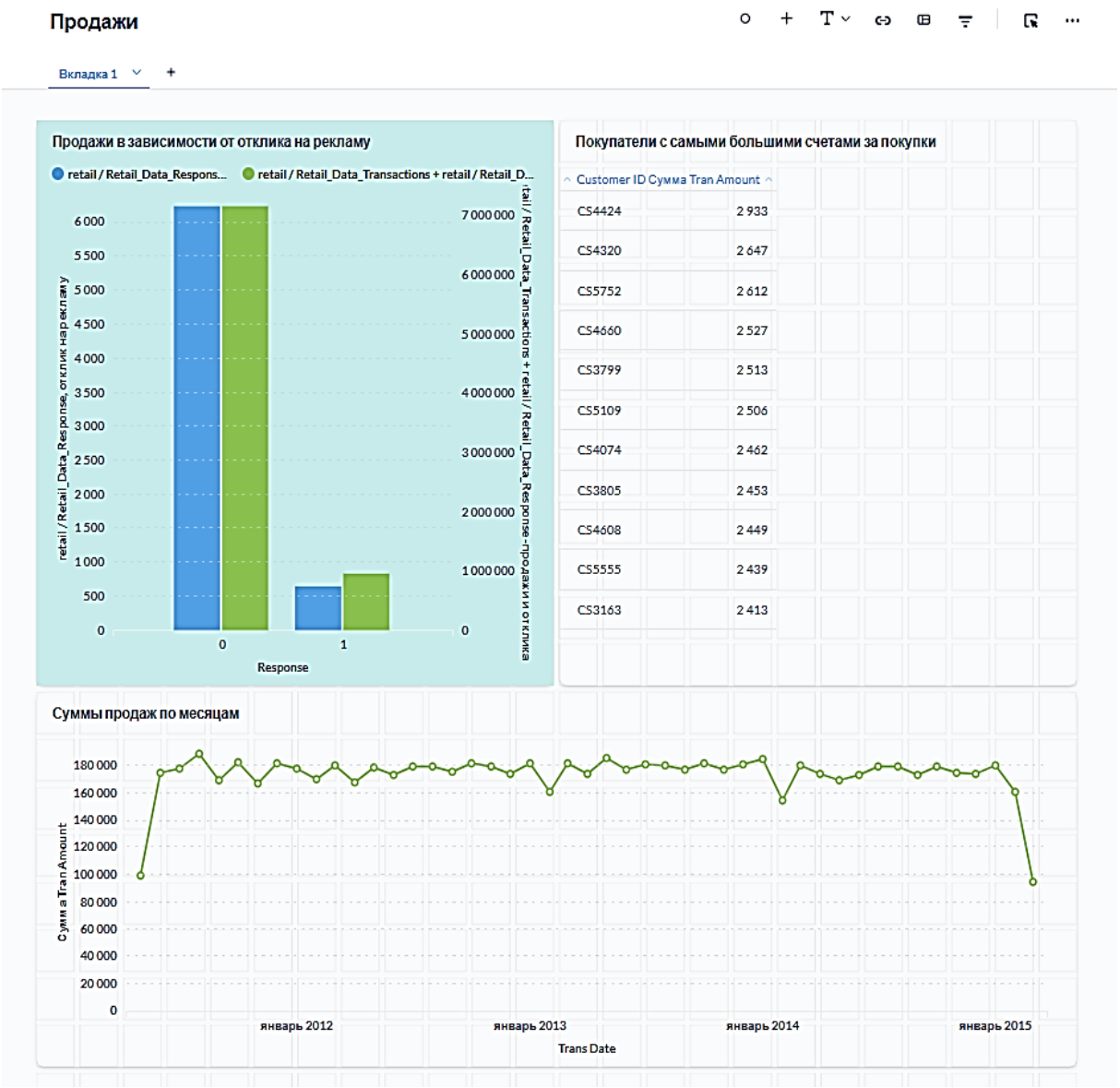


Рис. 1.70. Добавление новых визуализаций
на дашборд

Добавим фильтр по показателю trans_date (рис. 1.71–1.73).

Как видим, фильтр по времени не будет работать для объединенной диаграммы и для диаграммы по данным таблицы Retail_Data_Response.

Сохраним дашборд Продажи и внесем значения в фильтр, только одна из визуализаций изменится (рис. 1.75).

Для дальнейшего совершенствования навыков работы на платформе Glarus BI можно воспользоваться справочной документацией [2].

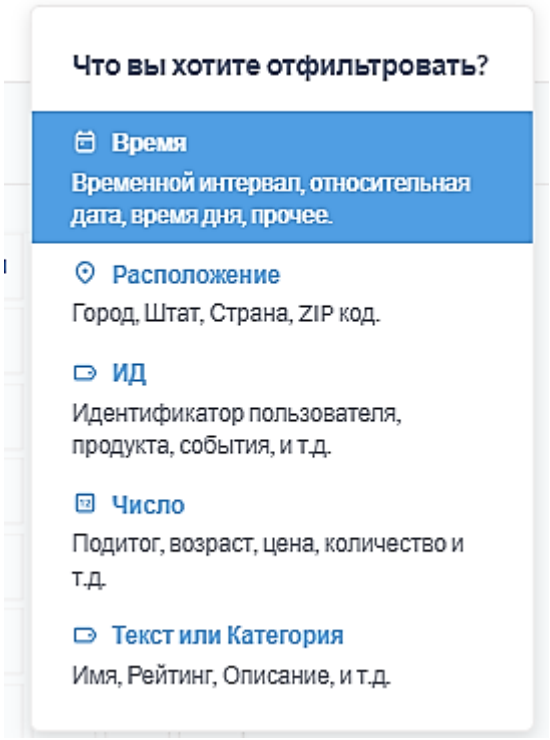


Рис. 1.71. Выбор типа показателя (поля) для настройки фильтра

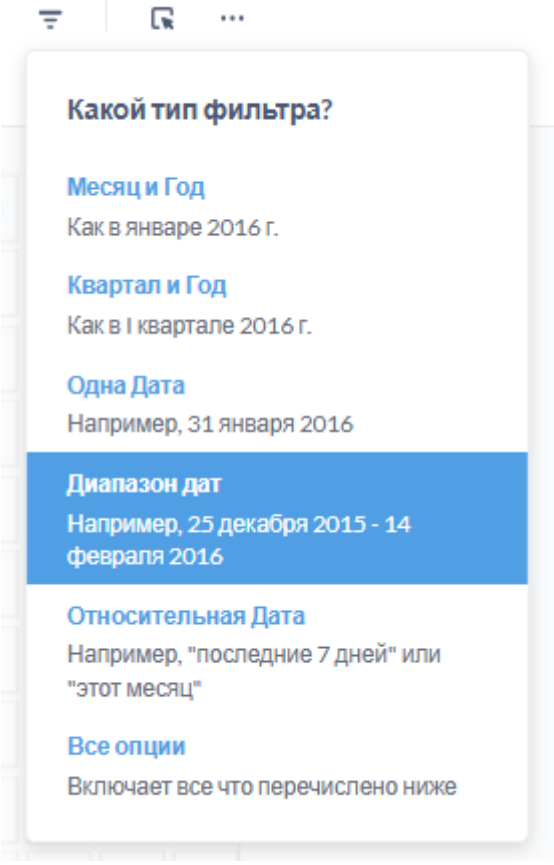


Рис. 1.72. Выбор типа фильтра

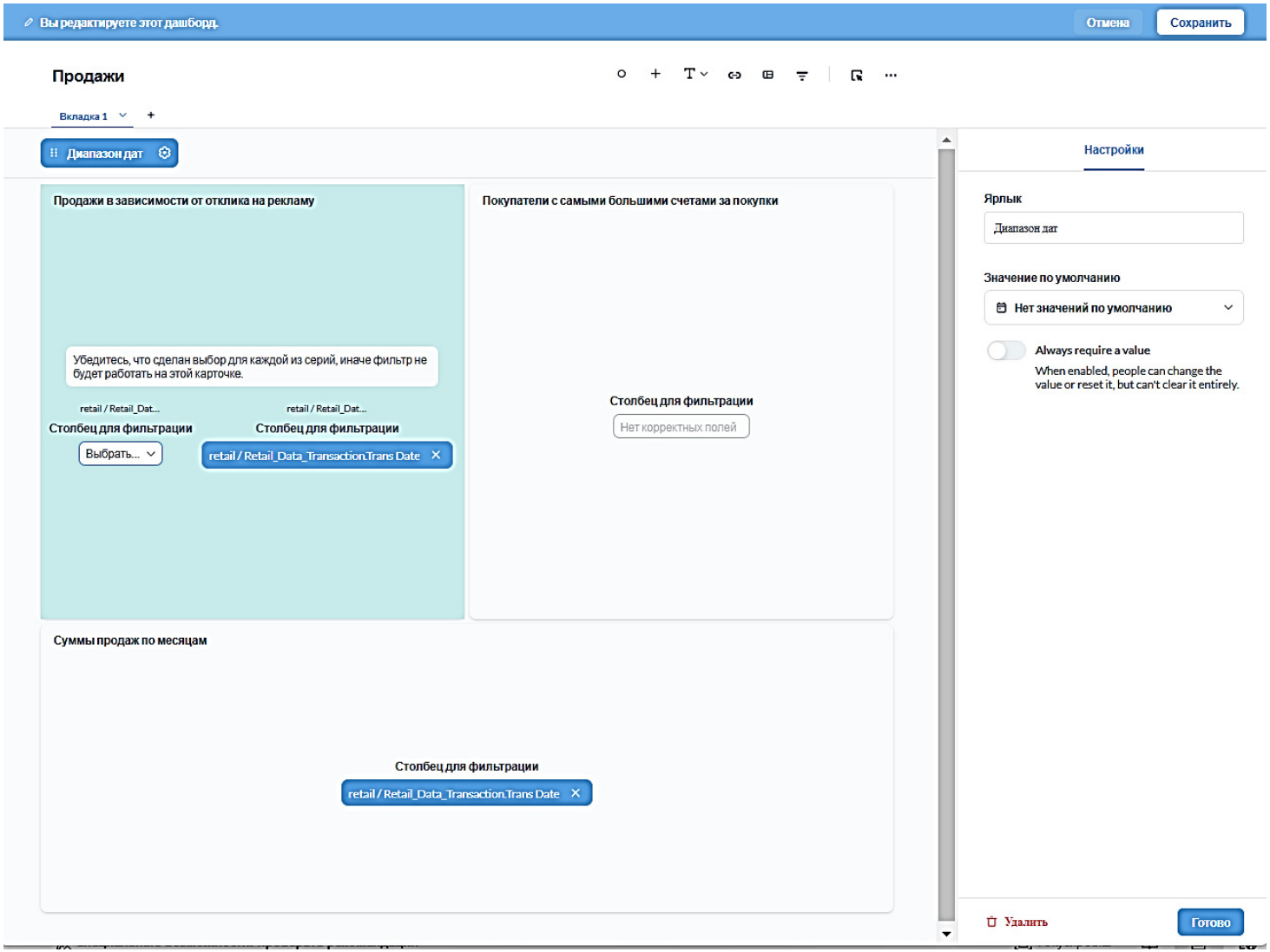


Рис. 1.73. Выбор показателя для фильтрации

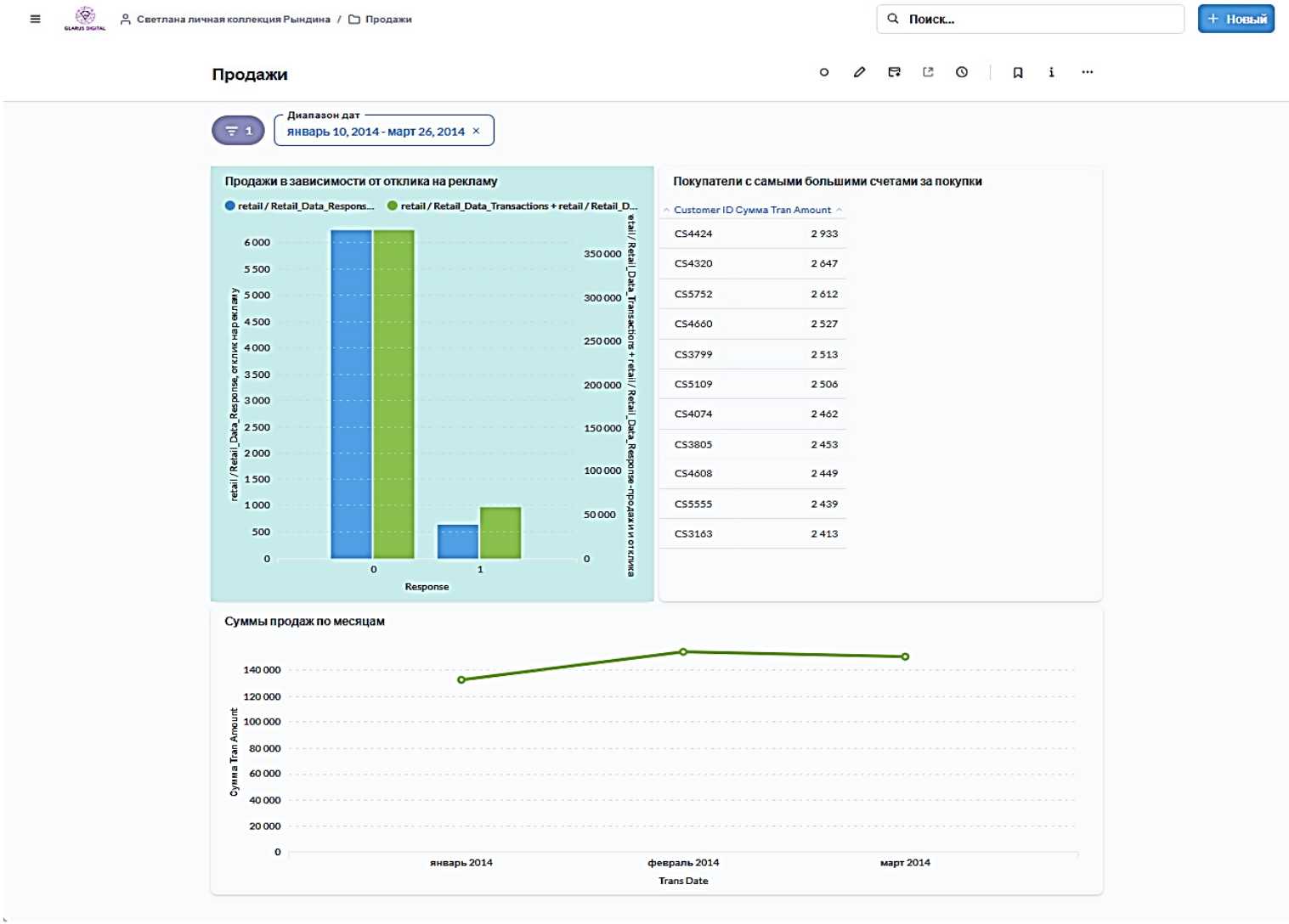


Рис. 1.74. Вид дашборда с примененным фильтром

2. РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ В YANDEX DATALENS

2.1. Yandex DataLens – система Business Intelligence

Yandex DataLens – это облачный сервис для бизнес-аналитики [3]. Сервис позволяет подключаться к различным источникам данных, строить визуализации, собирать дашборды и делиться полученными результатами.

С помощью Yandex DataLens можно отслеживать продуктовые и бизнес-метрики напрямую из источников, чтобы принимать решения, основанные на данных.

DataLens состоит из нескольких сущностей, которые обеспечивают полный цикл работы с данными (рис. 2.1):

– подключение – набор параметров для доступа к источнику данных;

– датасет (dataset, набор данных) – описание набора данных из источника;

– чарт (chart, диаграмма) – визуализация данных из датасета в виде таблиц, диаграмм и карт;

– дашборд (dashboard, приборная панель) – набор чартов, селекторов для фильтрации данных и текстовых блоков.

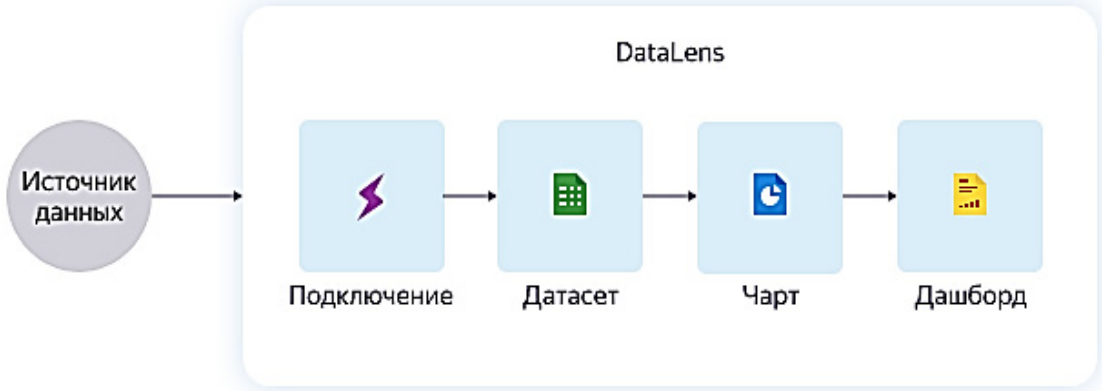


Рис. 2.1. Сущности DataLens

В сервис можно зайти по адресу: https://datalens.yandex.ru (рис. 2.2) с учетной записью Yandex (аккаунт на @yandex.ru).

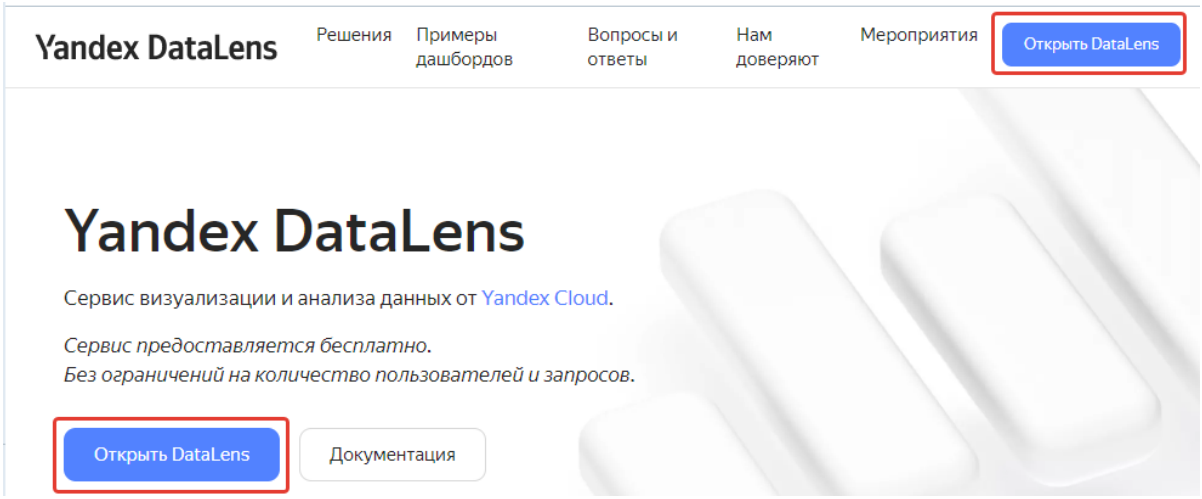


Рис. 2.2. Страница для входа в сервис

На стартовой странице (перейти на которую можно по виджету в левом верхнем углу) доступны для создания все сущности (Подключения, Датасеты, Чарты и Дашборды). Слева располагается вертикальная панель навигации (рис. 2.3).

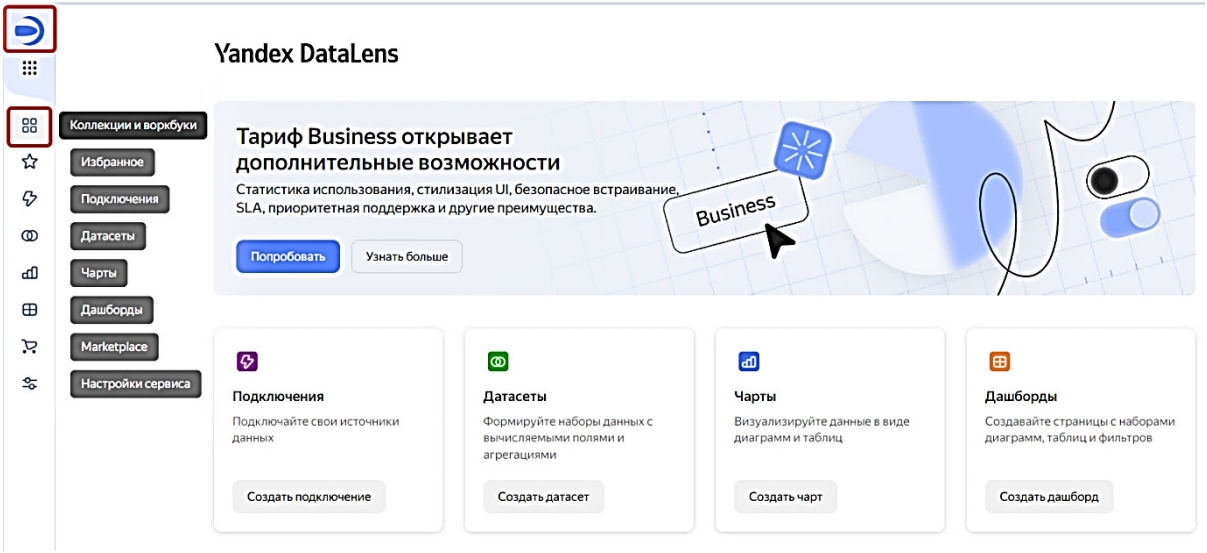


Рис. 2.3. Стартовая страница DataLens

Сущность Подключение содержит информацию о параметрах доступа к источнику данных, на базе подключения можно на следующем шаге создавать Датасеты.

В DataLens доступно множество типов подключений (рис. 2.4–2.6).

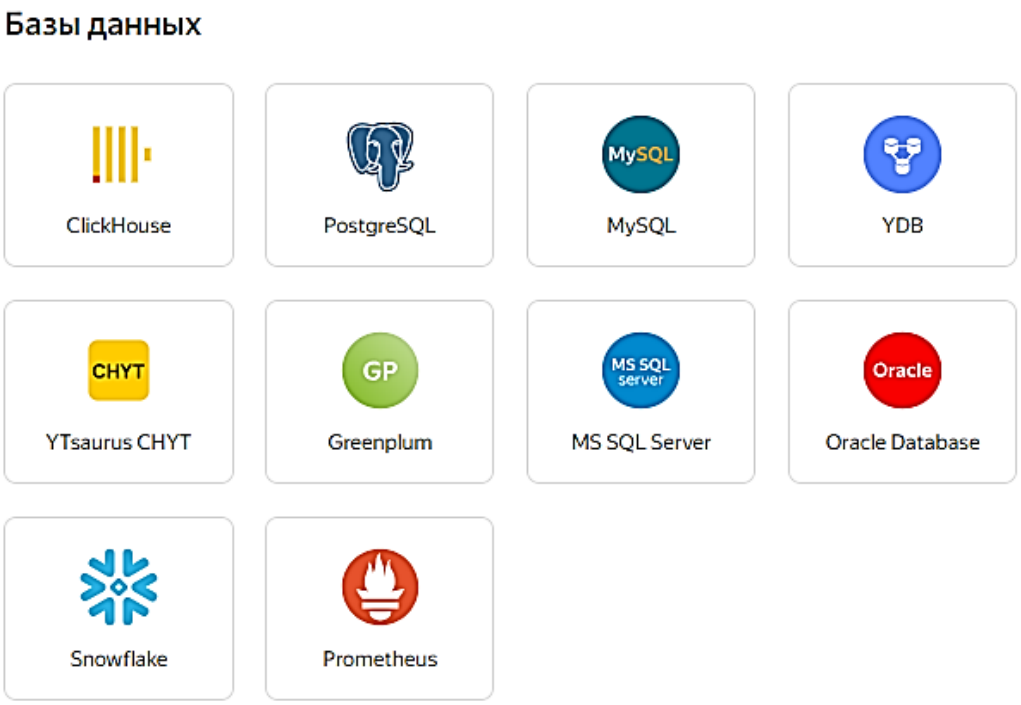


Рис. 2.4. Источники данных (базы данных)

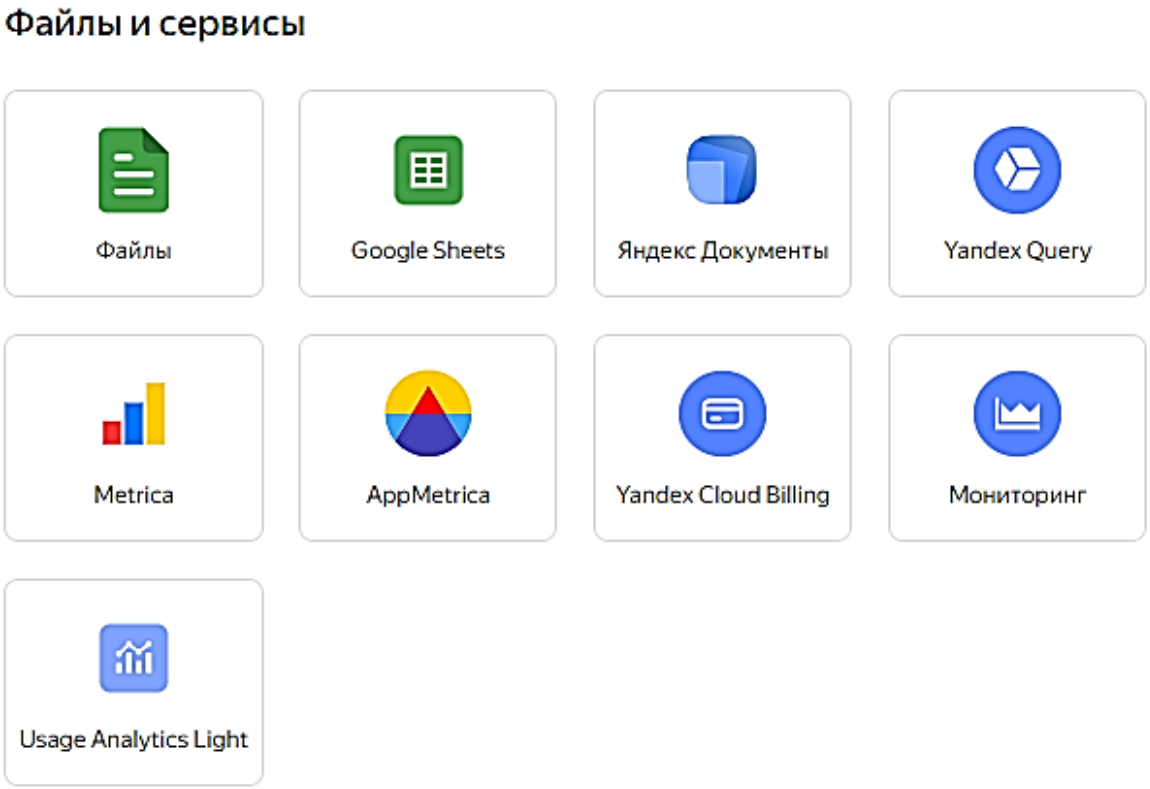


Рис. 2.5. Источники данных (файлы и сервисы)

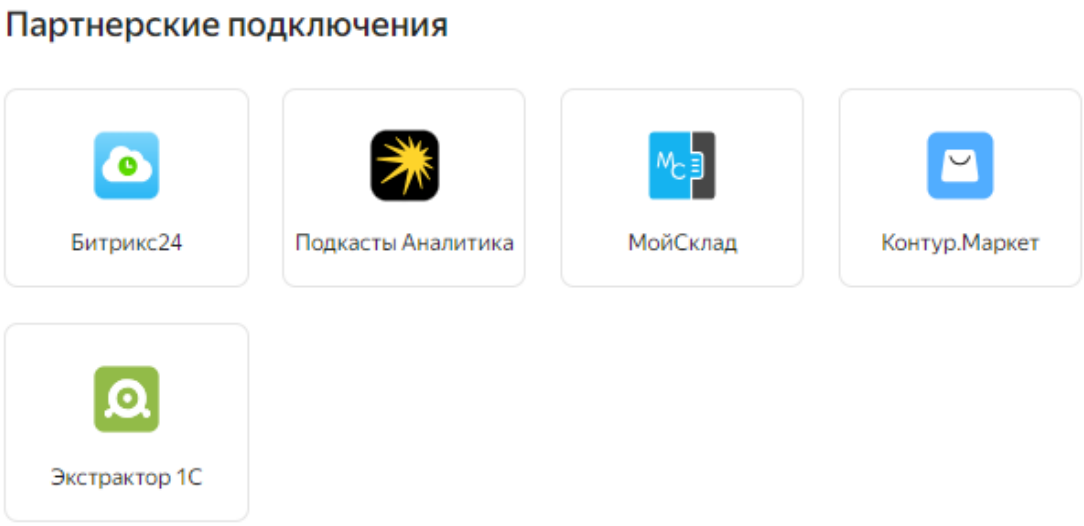


Рис. 2.6. Источники данных (партнерские подключения)

Файлы с данными могут иметь расширение .xlsx и .csv. Большое количество наборов данных доступно для создания отчетов и проведения анализа именно в формате .csv. Это текстовый формат для представления табличных данных (Comma-Separated Values, т.е. значения, разделенные запятыми). Многие программы для работы с табличными данными, в том числе и реляционные системы управления данными (PostgreSQL, MySQL и др.), позволяют импортировать данные в этом формате.

2.2. Создание подключения и датасета

Выберем на главной странице опцию Создать подключение и перейдем к варианту Файлы. В диалоговом окне появится возможность загрузить файл в облако, а при его изменении придется создавать новое подключение. Также нельзя превышать предельный размер файла – 200 Мб (с учетом бесплатного использования сервиса).

Загрузим файл insurance.csv, с которым мы работали в предыдущем разделе (рис. 2.7). Переводить его в формат .xlsx не требуется. При считывании содержимого файла нужно контролировать кодировку и разделитель (они распознаются автоматически), а также правильность определения типов данных: числовых с маркером # и текстовых с маркером T.

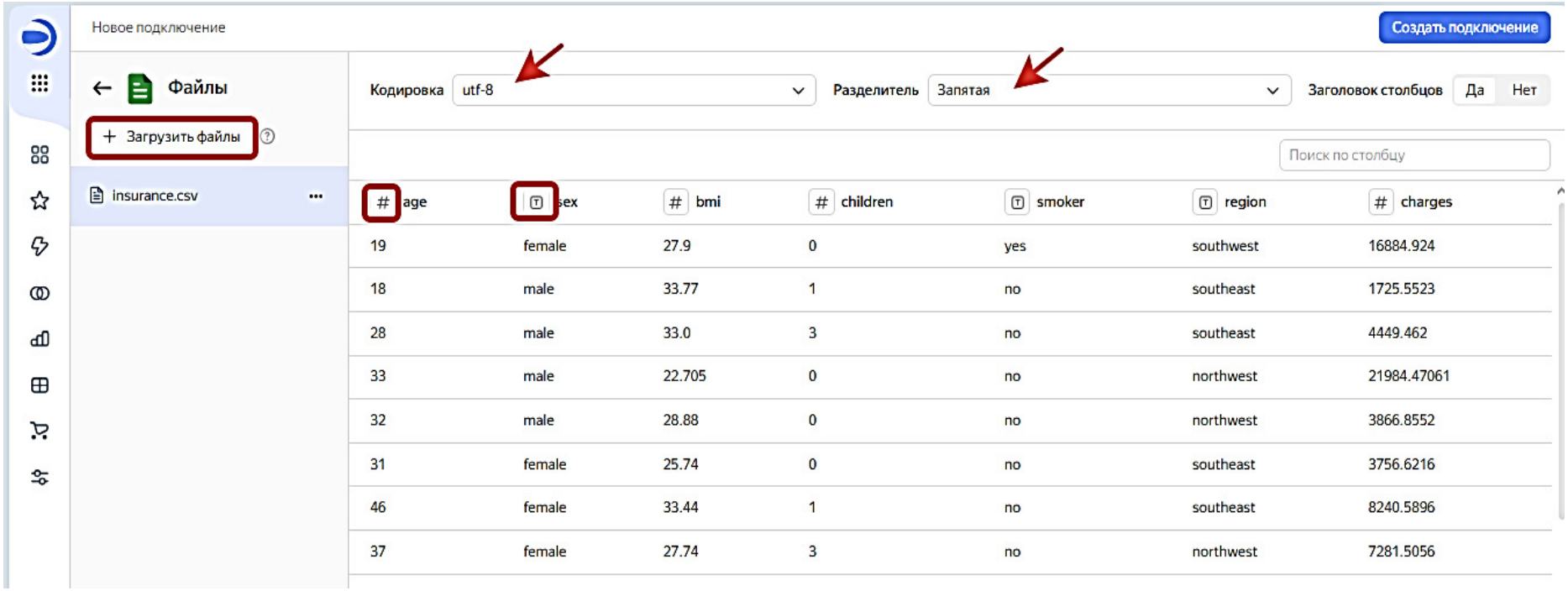


Рис. 2.7. Загрузка файла

Для сохранения подключения необходимо создать воркбук (рис. 2.8).

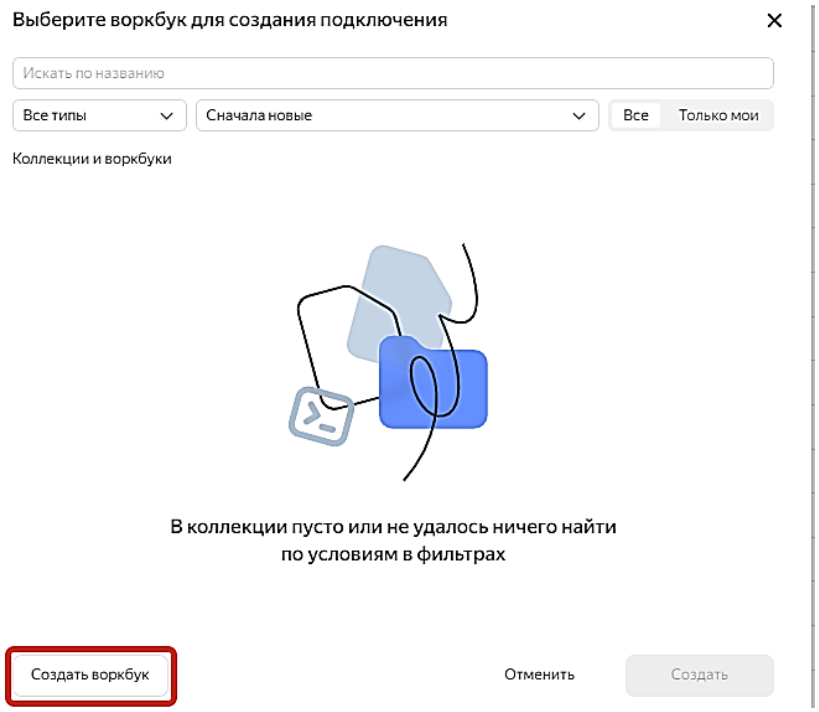


Рис. 2.8. Создание воркбука

Присвоим ему имя Медицинское страхование и сможем сохранить в этот воркбук подключение с именем insurance (рис. 2.9).

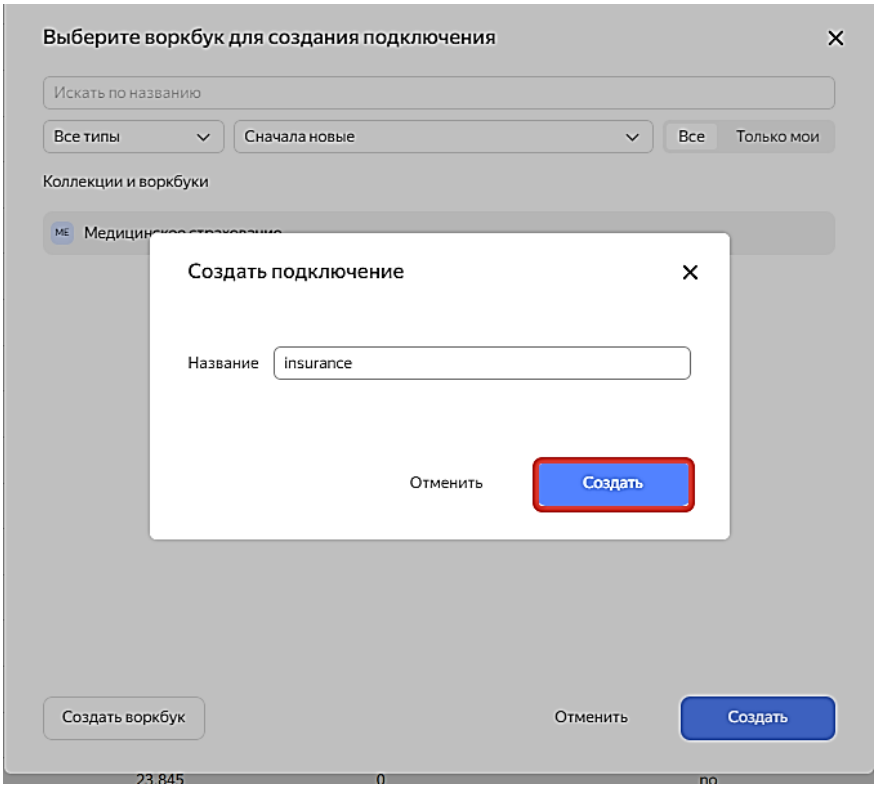


Рис. 2.9. Сохранение подключения

После сохранения подключения становится доступным создание датасета (рис. 2.10).

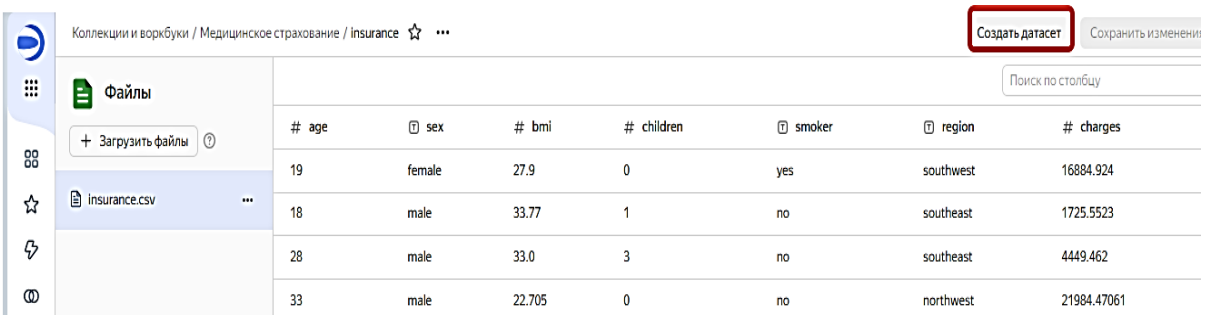


Рис. 2.10. Создание датасета

При создании датасета можно произвести необходимые настройки (создание датасета в чем-то похоже на запрос в Glarus BI): скрыть поле, настроить агрегацию, добавить описание и т.п. (рис. 2.11).

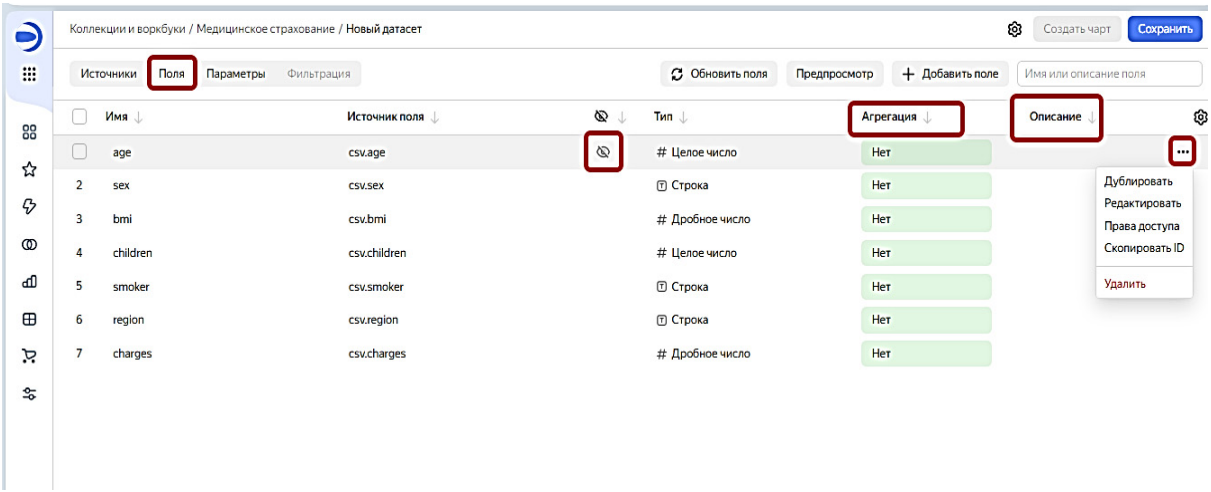


Рис. 2.11. Настройки для датасета

Поля определяют структуру и формат датасета. Существуют следующие типы полей:

– Измерение. Такие поля используются для разделения наборов данных на группы. Например: город, дата покупки, категория продукта. Для такого поля определены следующие агрегации (aggregatio, присоединение – это процесс объединения элементов): количество значений в поле и количество уникальных значений (рис. 2.12).

– Показатель. Содержит числовые значения, к которым применяются функции агрегации (сведения). Например: сумма кликов, количество переходов (рис. 2.13).

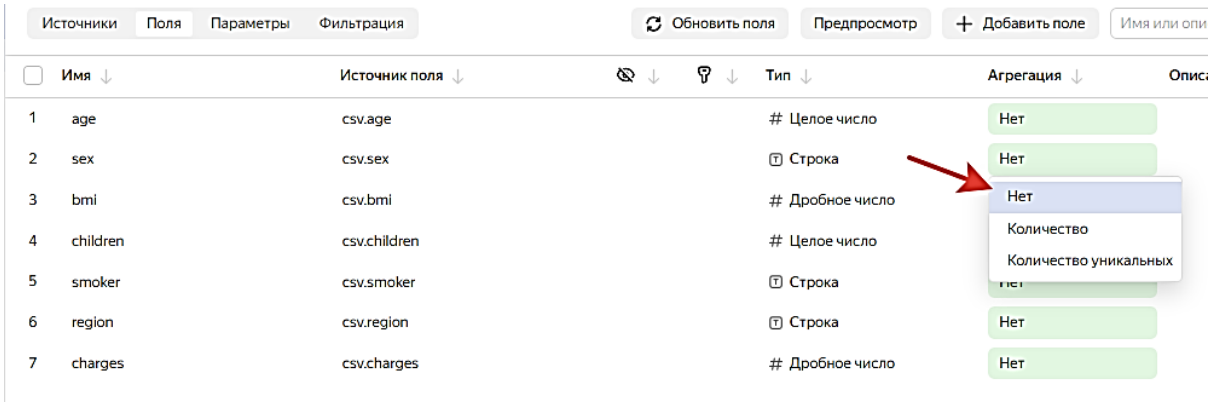


Рис. 2.12. Варианты агрегации для измерений

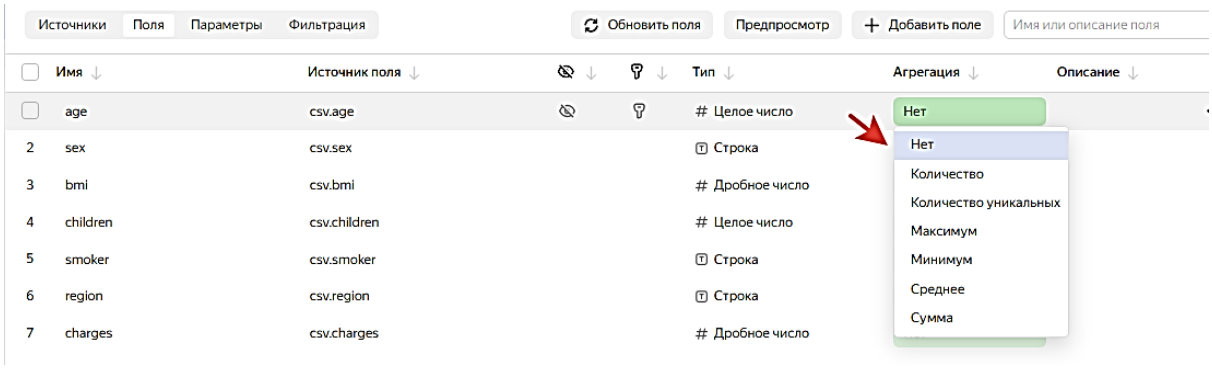


Рис. 2.13. Варианты агрегации для показателей

Для одного подключения можно создать несколько датасетов. Однако в этом примере будем работать с исходной таблицей. Сохраним датасет с тем же именем insurance.

2.3. Создание чартов

В сохраненном датасете появляется опция Создать чарт (рис. 2.14).

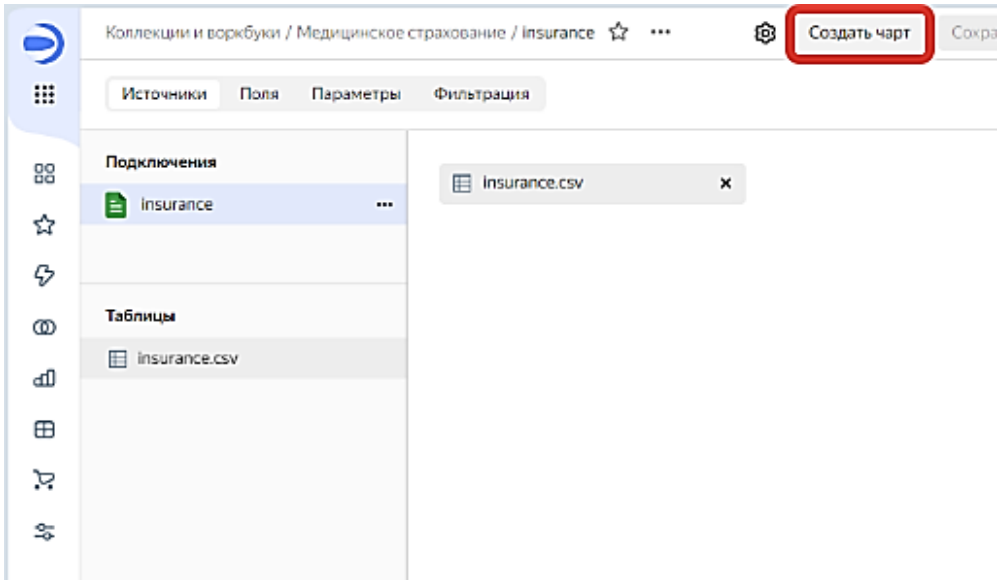


Рис. 2.14. Создание чарта

Создадим диаграмму разброса расходов от веса. Изменим тип чарта на точечную диаграмму (рис. 2.15).

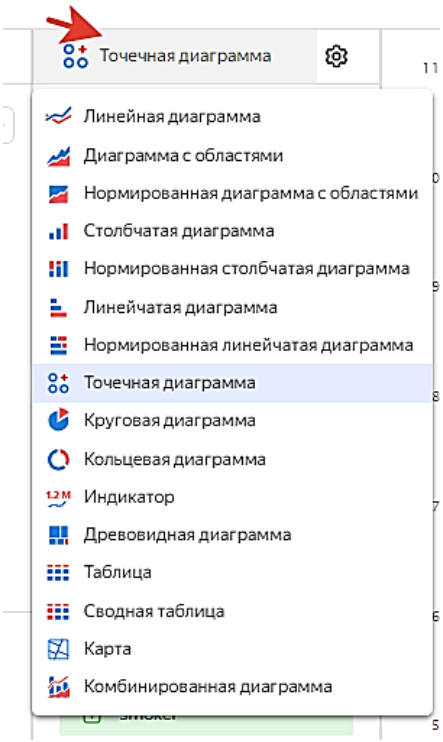


Рис. 2.15. Выбор типа диаграммы

Добавим в качестве оси X показатель bmi, в качестве оси Y показатель charges, цвет точек определим с помощью показателя smoker, а форму точек с помощью показателя sex (рис. 2.16). При наведении на точку графика курсора появляются значения показателей для нее.

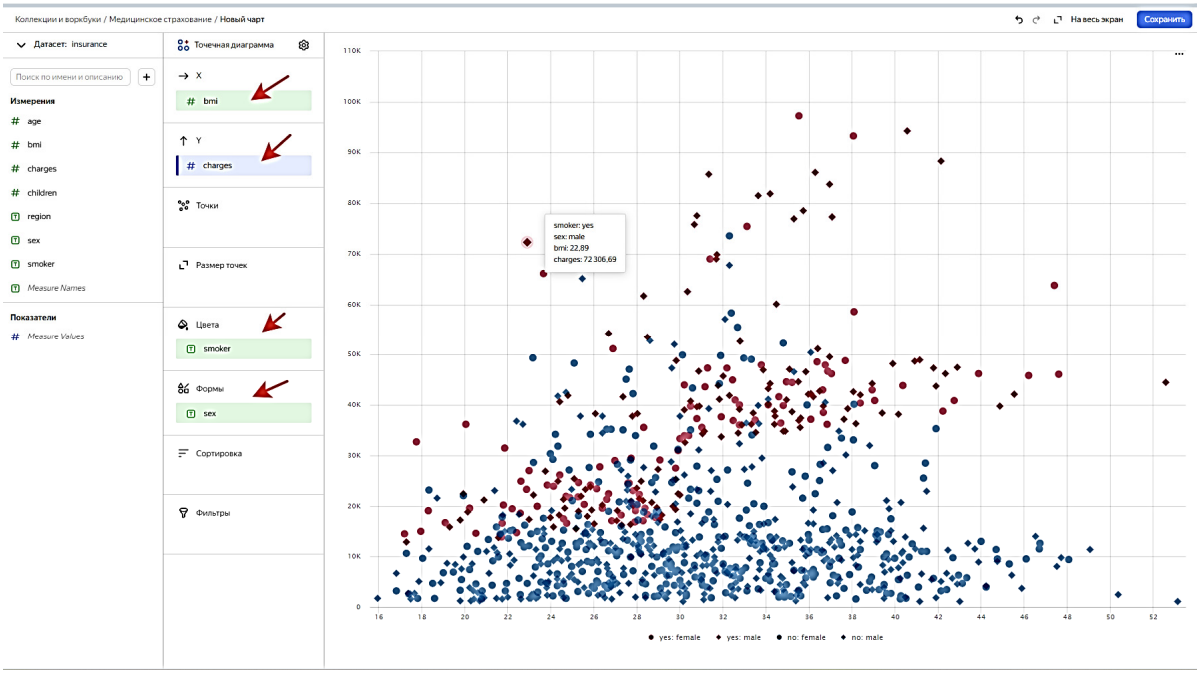


Рис. 2.16. Настройки для диаграммы

При наведении курсора на элемент в панели настройки чарта появляется знак настроек ⚙ для конкретной опции (всего чарта, его осей, цвета и т.п.). Так, например, можно добавить на диаграмму заголовок (рис. 2.17).

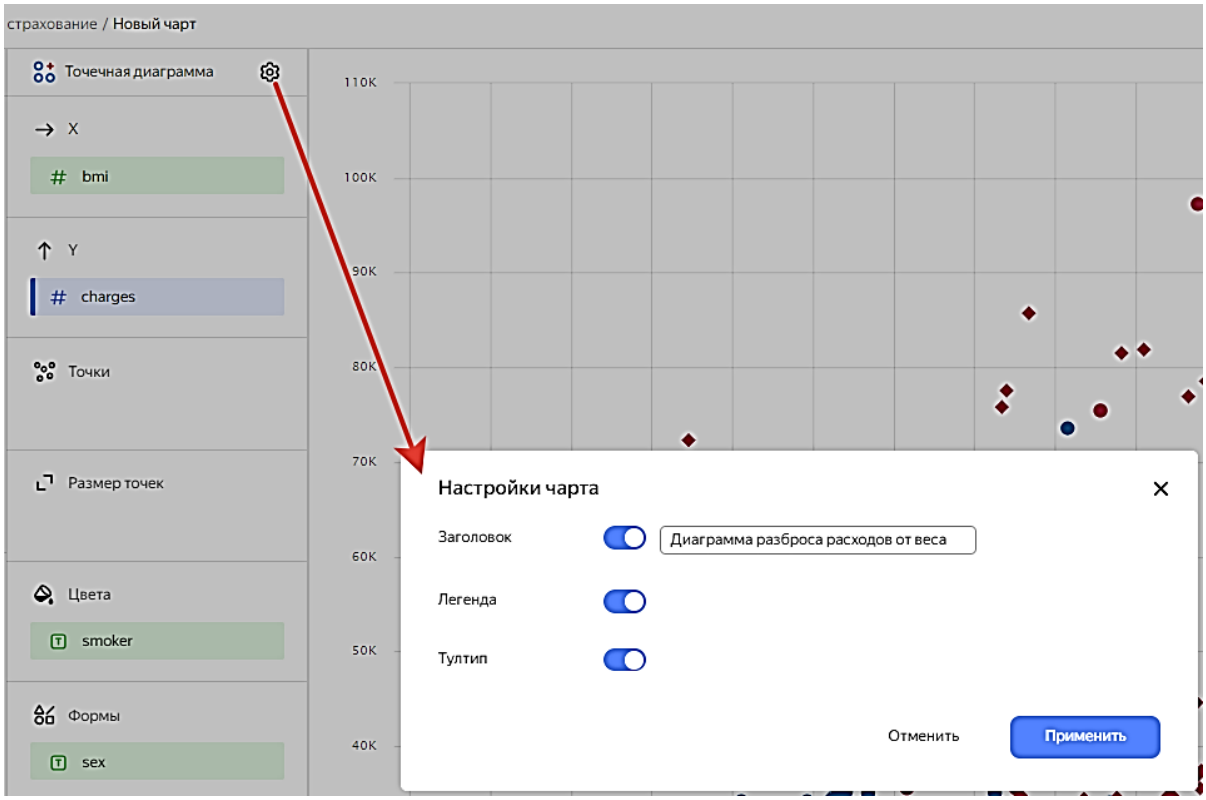


Рис. 2.17. Добавление заголовка

Можно изменить настройки для формы точек на графике (рис. 2.18).

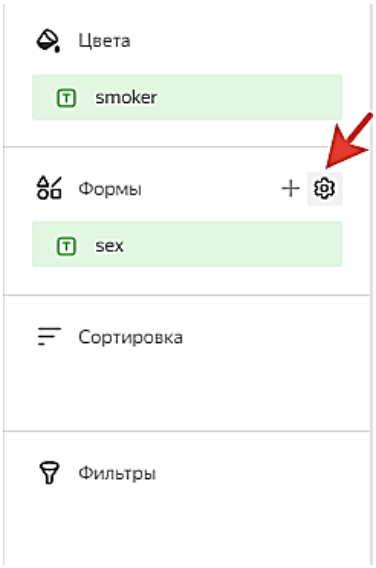
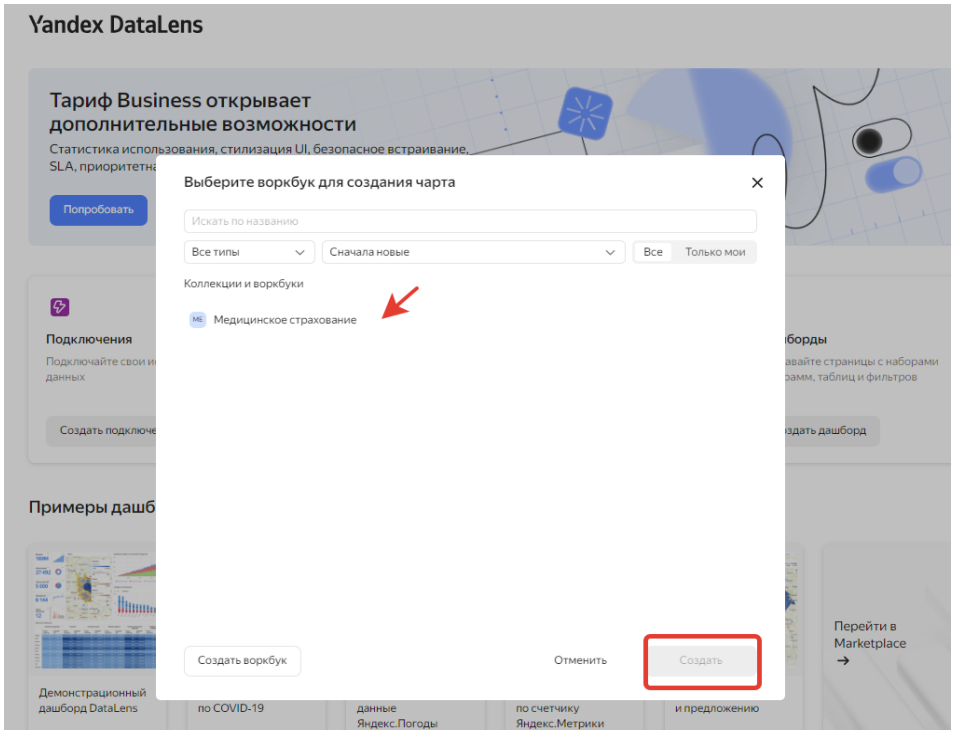


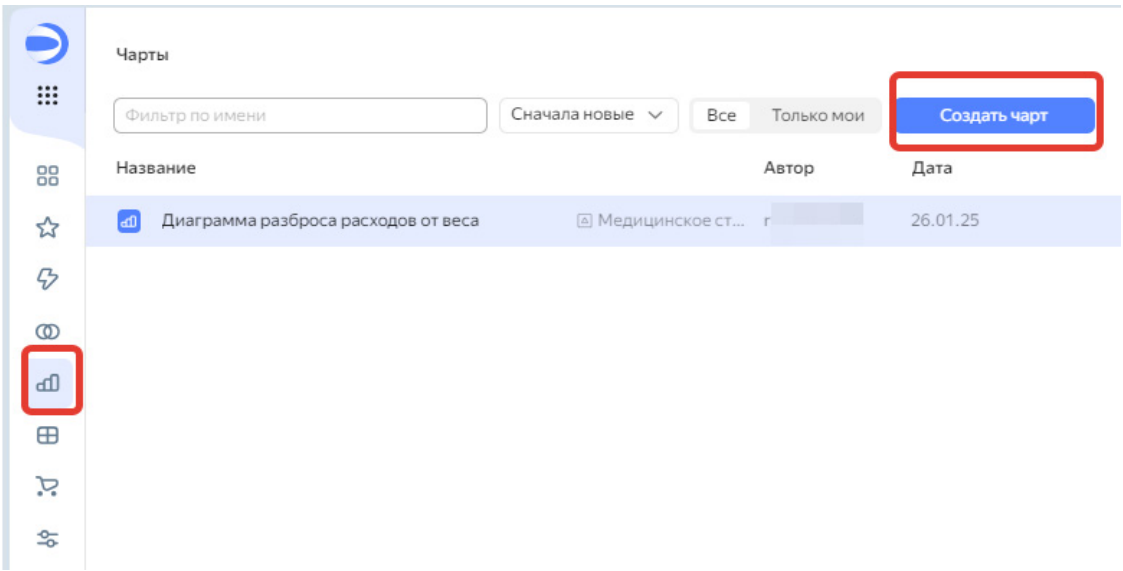
Рис. 2.18. Настройки формы точек на графике

Размер точек поменяем на 12, чтобы сделать визуализацию более информативной.

Создадим новый чарт. Это можно сделать различными способами: перейти на главную страницу сайта и выбрать виджет – Создать (рис. 2.19,*а*), можно в пиктографическом меню, расположенном слева, кликнуть на кнопке чарты и в открывшейся вкладке нажать кнопку Создать чарт (рис. 2.19,*б*).



а)



б)

Рис. 2.19. Создание чарта с помощью виджета (*а*)
и с помощью пиктографического меню (*б*)

В этом случае необходимо будет выбрать данные для построения чарта (рис. 2.20).

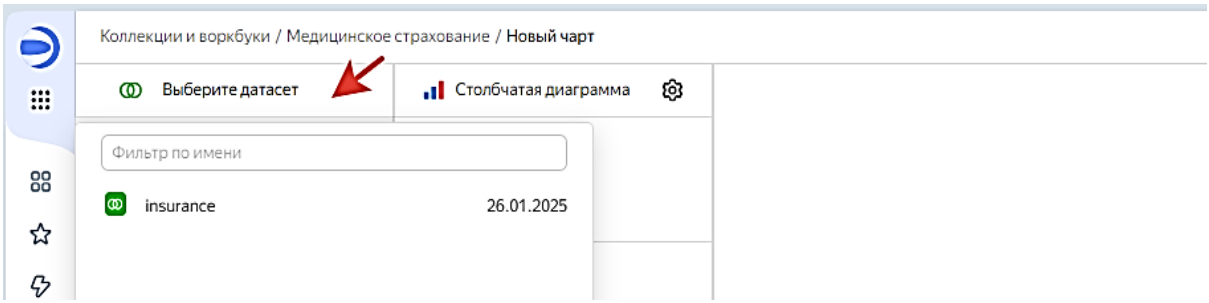


Рис. 2.20. Выбор датасета

Можно открыть датасет, использовав пиктографическое меню (рис. 2.21) и создать чарт из датасета (см. рис. 2.15).

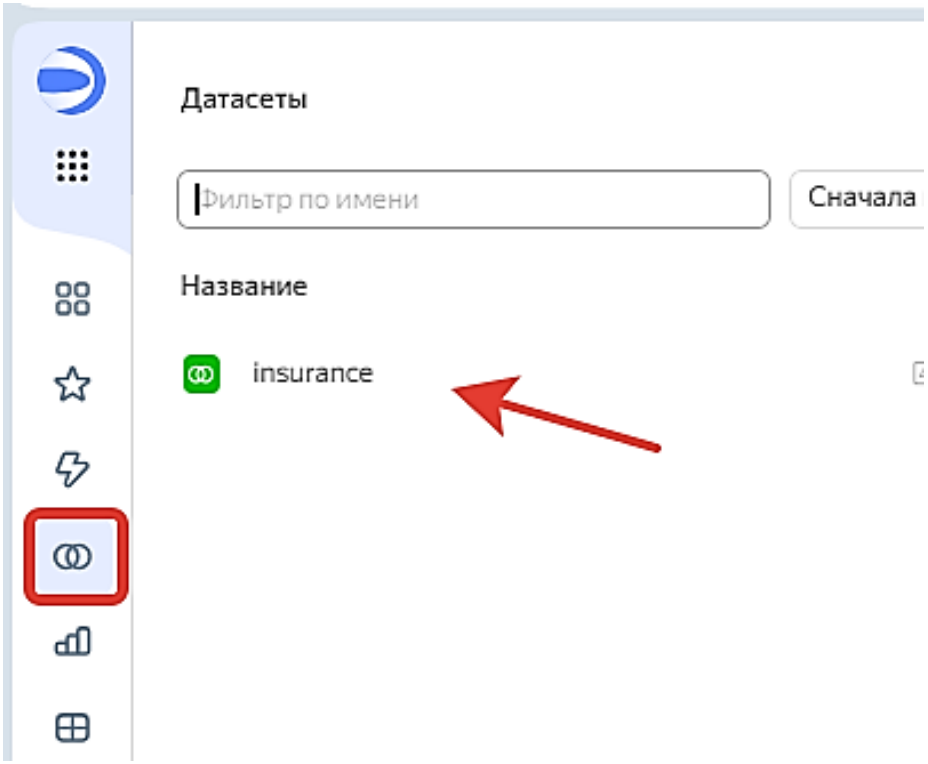


Рис. 2.21. Обращение к датасету

В чарте, открытом для настройки, построим столбчатую диаграмму по переменной region, чтобы визуализировать число наблюдений, соответствующих каждому ее значению.

По оси X – значения категориальной переменной region, а по оси Y – агрегация для этого показателя, чтобы сменить агрегацию, нужно кликнуть по маркеру # слева от имени показателя (рис. 2.22).

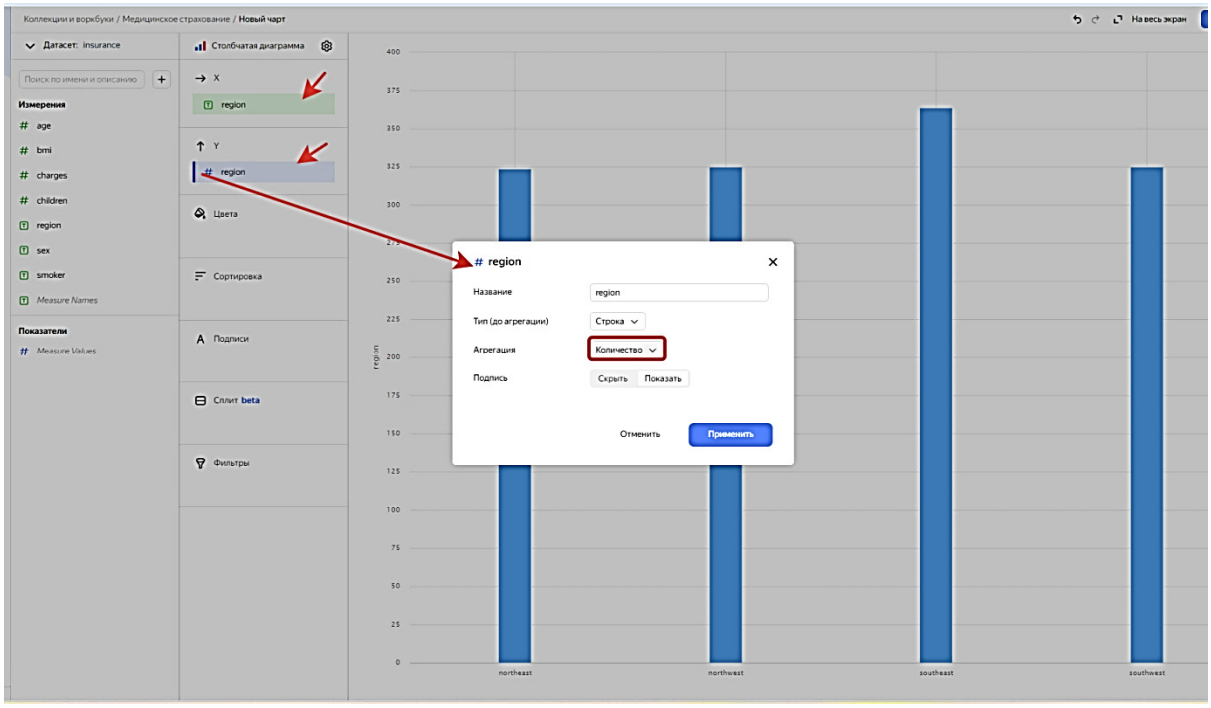


Рис. 2.22. Настройки для чарта

Настроим заголовок Число наблюдений для каждого значения переменной region (рис. 2.23). С тем же именем сохранится чарт.

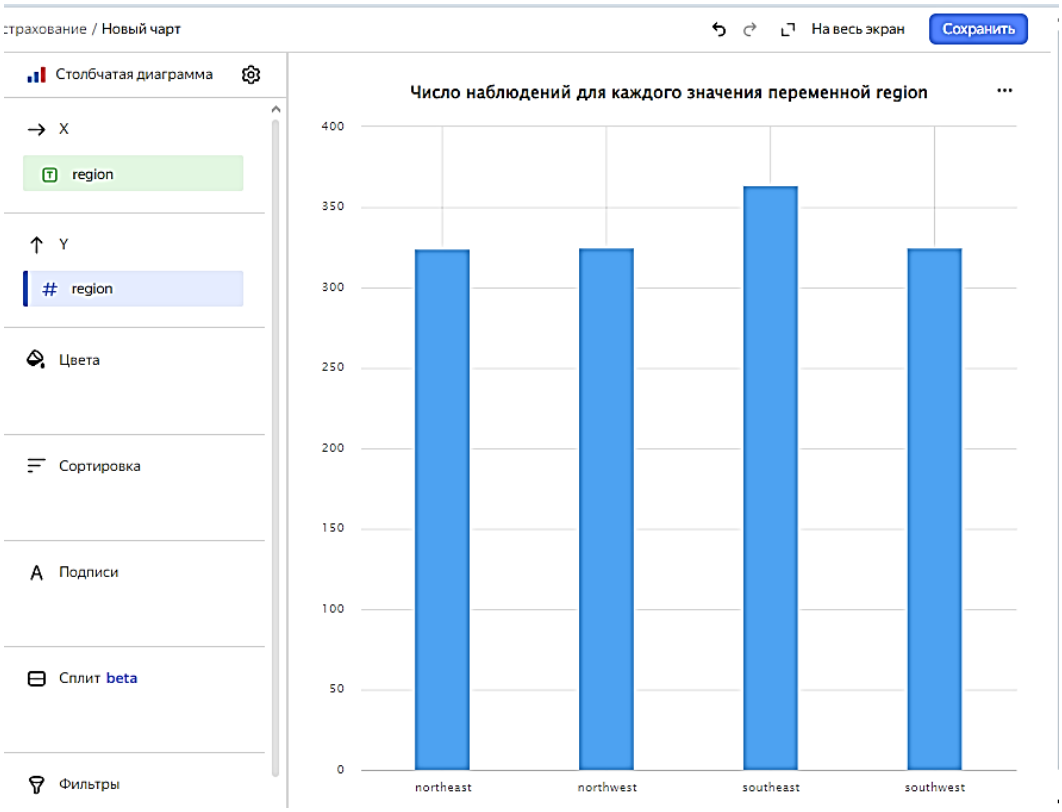


Рис. 2.23. Добавление заголовка

Построим на основе этого чарта новый, в который добавим разбиение наблюдений в каждом из районов по показателю smoker, отвечающему за наличие вредной привычки, изменим заголовок и сохраним как копию (рис. 2.24). Однако лучше все же сначала сохранить первоначальный чарт как копию, а только потом вносить изменения.

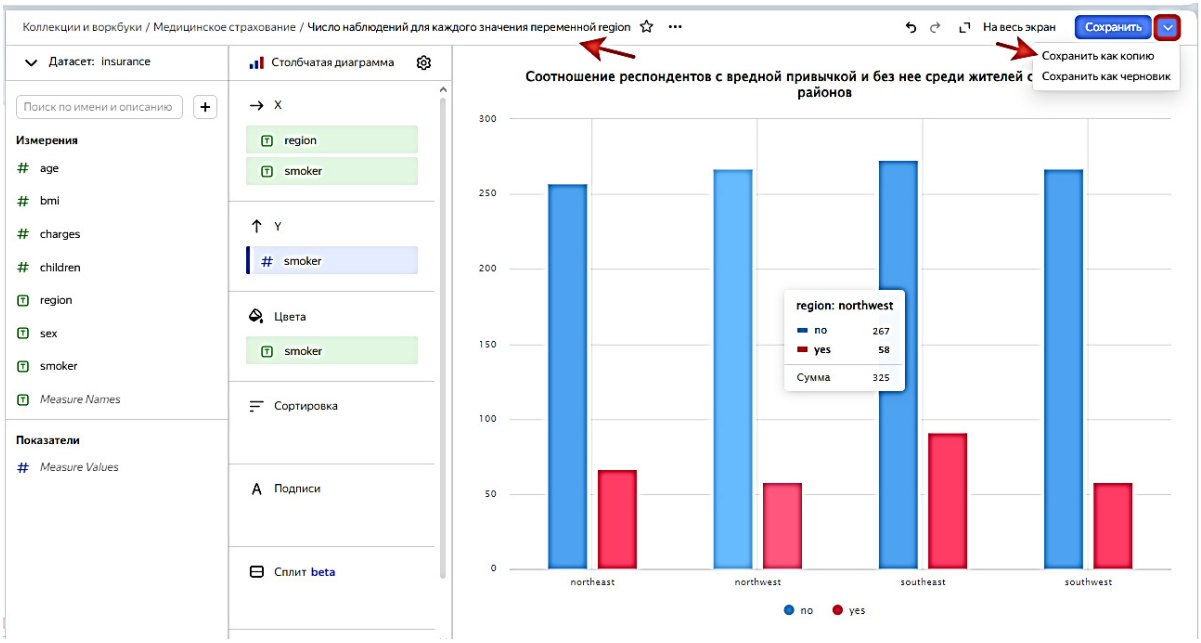


Рис. 2.24. Новый чарт на основе ранее созданного

Для чарта доступны дополнительные опции, которые позволяют просмотреть таблицу с данными, сохранить чарт в файл и т.д. (рис. 2.25).

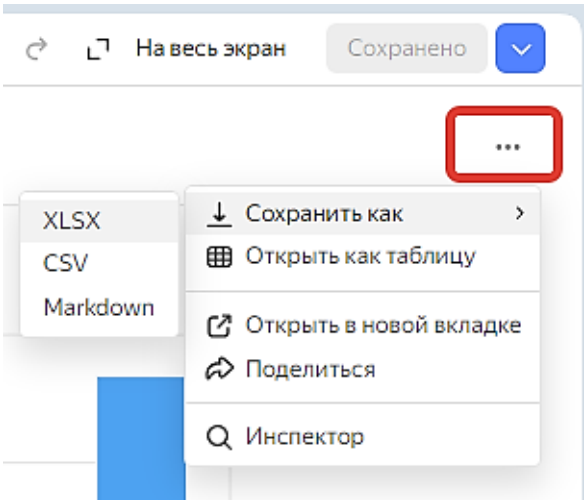


Рис. 2.25. Дополнительные возможности для работы с чартом

Приступим к созданию дашборда.

2.4. Создание дашборда

Перейдем на главную страницу сервиса и выберем опцию Создать дашборд (см. рис. 2.3). При создании необходимо выбрать воркбук, в котором он будет размещен в момент сохранения.

Дашборд состоит из элементов, называемых виджетами (widget приспособление), которые показывают некоторую информацию или выполняют несложное действие.

Yandex DataLens содержит следующие типы виджетов (рис. 2.26):

– чарт;
– селектор;
– текст;
– заголовок.

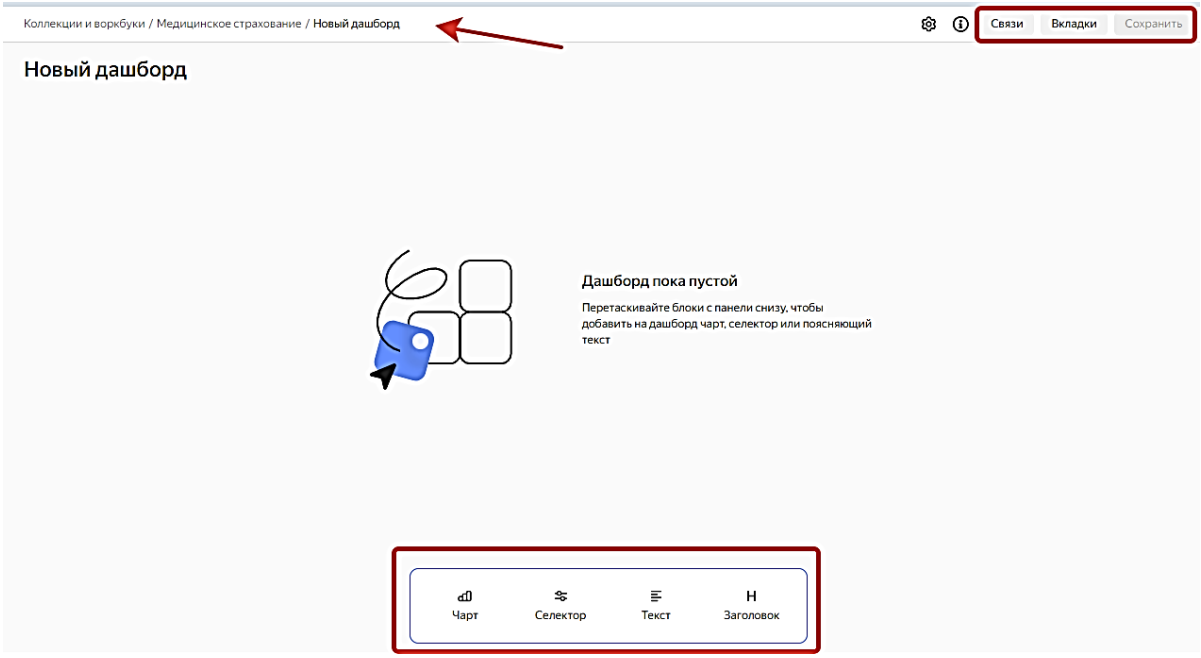


Рис. 2.26. Рабочая область дашборда

Виджет в виде текста позволяет разместить на дашборде пояснение, ссылки и другие элементы. Заголовки помогают структурировать виджеты внутри дашборда, отделяя группы связанных виджетов визуально.

Ранее созданные чарты можно размещать в любом месте страницы дашборда и изменять размеры. Селектор – это фильтр, который влияет на результаты запросов на связанных с ним виджетах (чарты или другие селекторы).

Добавим на дашборд ранее созданные чарты (рис. 2.27, 2.28).

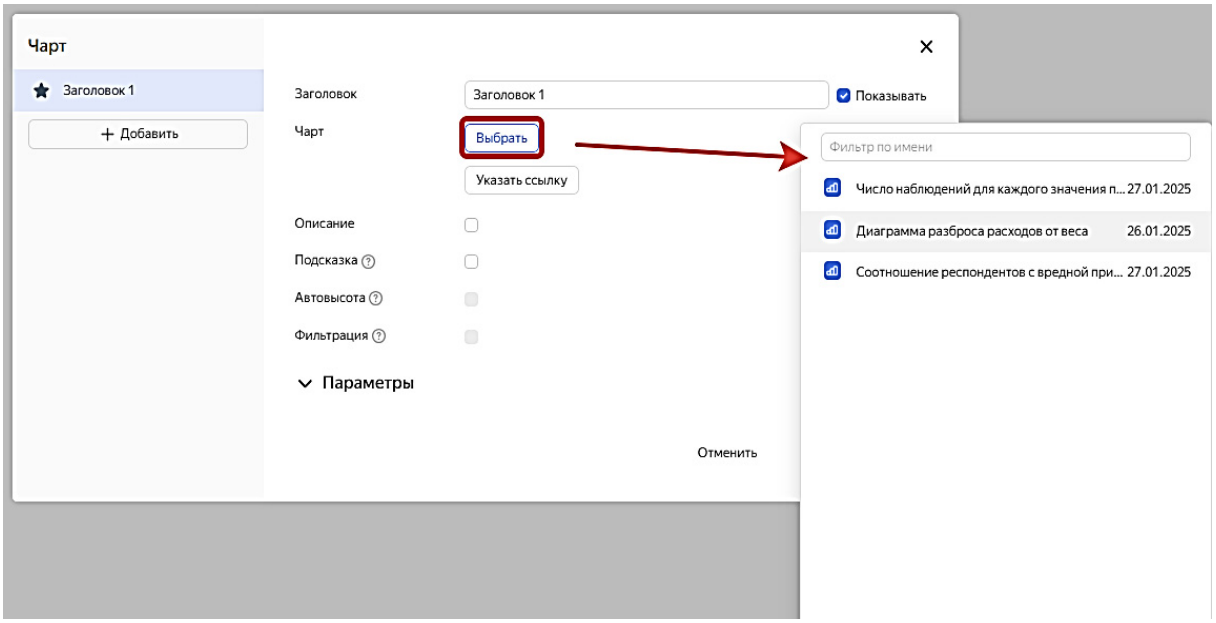


Рис. 2.27. Выбор чарта из ранее созданных

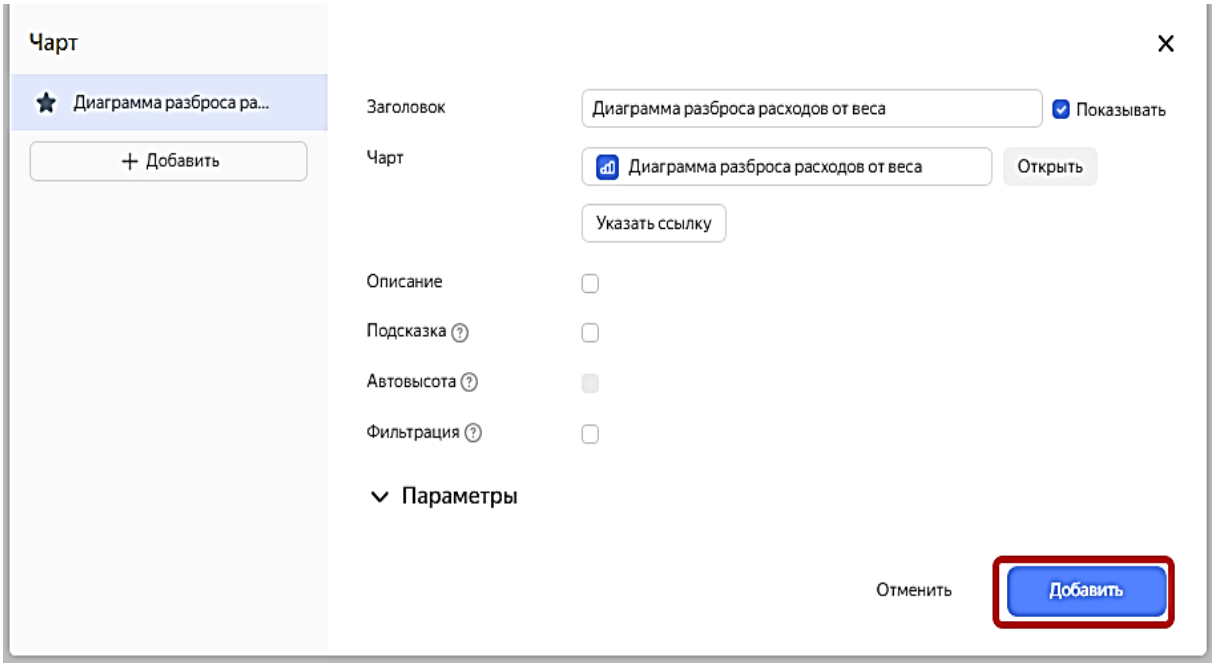


Рис. 2.28. Добавление чарта

Уберем заголовок добавленного чарта, перейдя в настройки виджета (рис. 2.29).

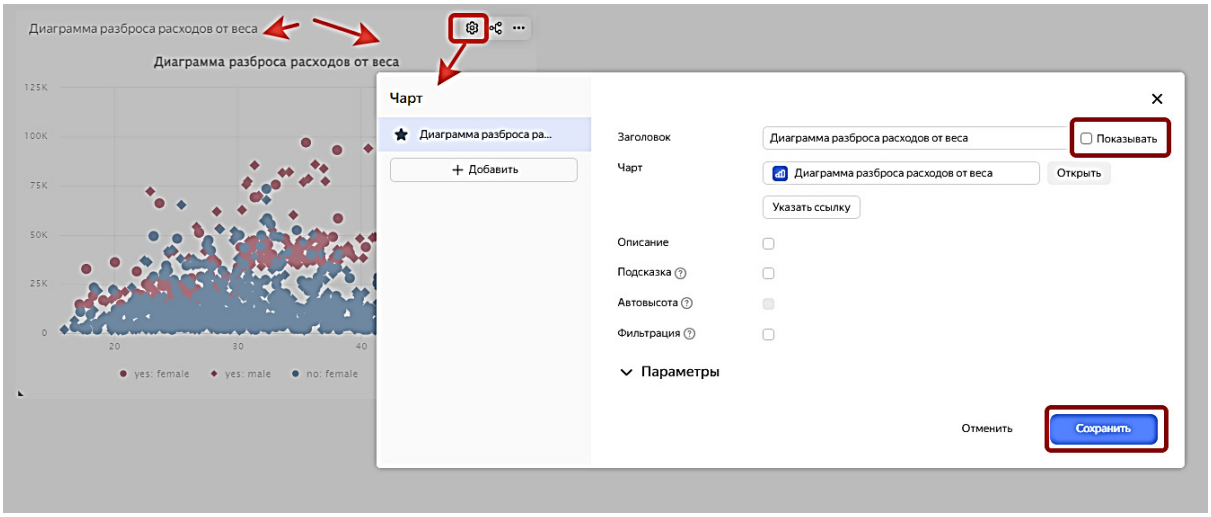


Рис. 2.29. Скрытие заголовка чарта с экрана

Добавим также остальные чарты на дашборд. Создадим фильтр по показателю sex (рис. 2.30, 2.31).

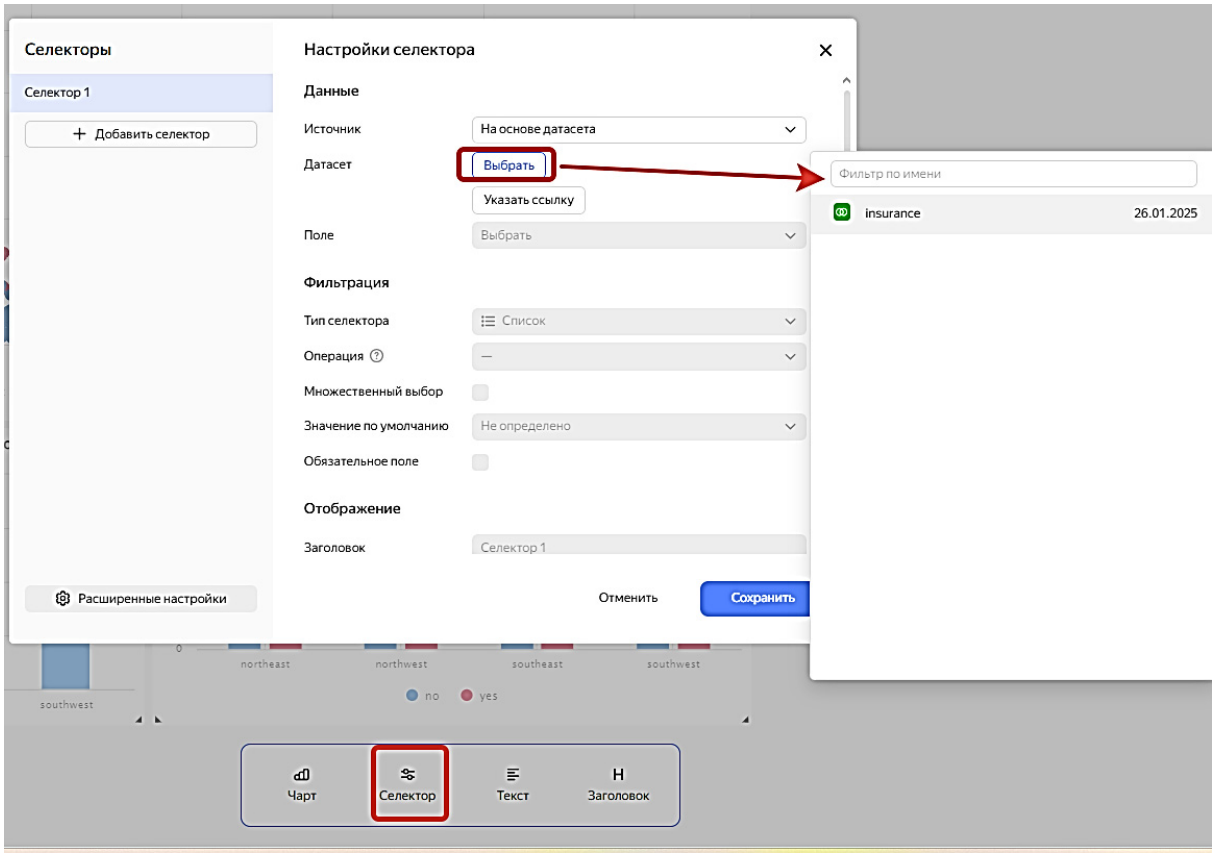


Рис. 2.30. Добавление селектора
и выбор набора данных

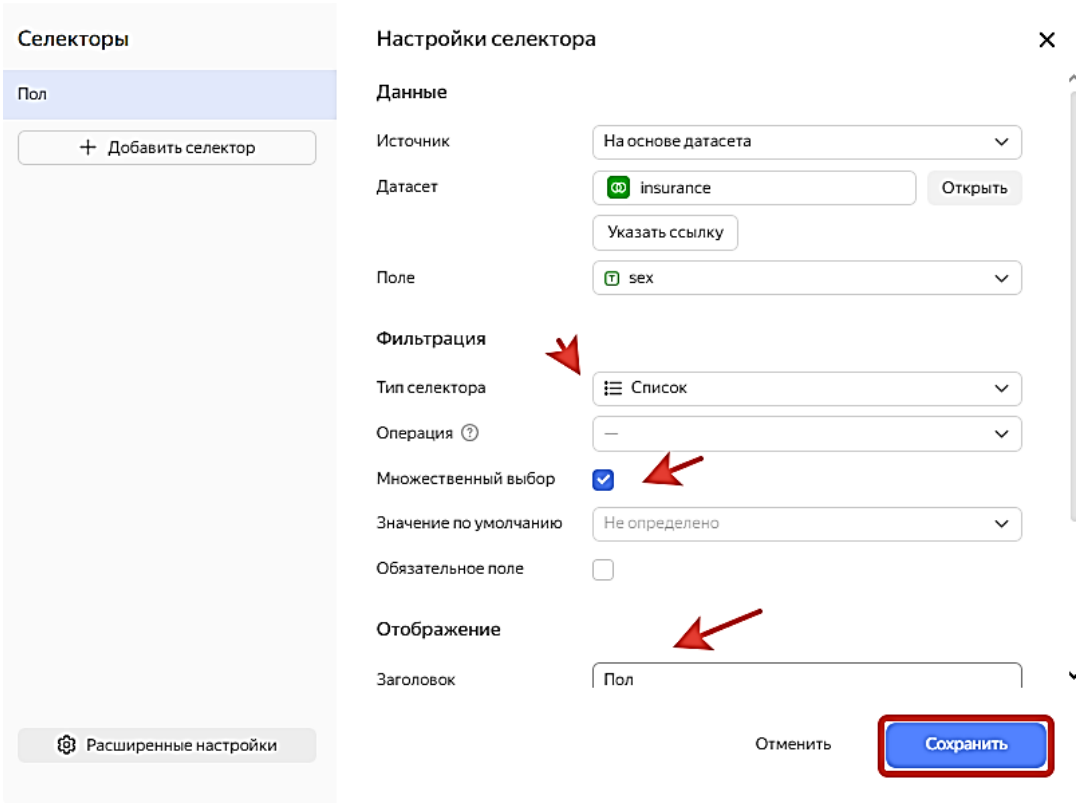


Рис. 2.31. Настройки для селектора по полу

Также добавим селектор по числовому показателю age (рис. 2.32).

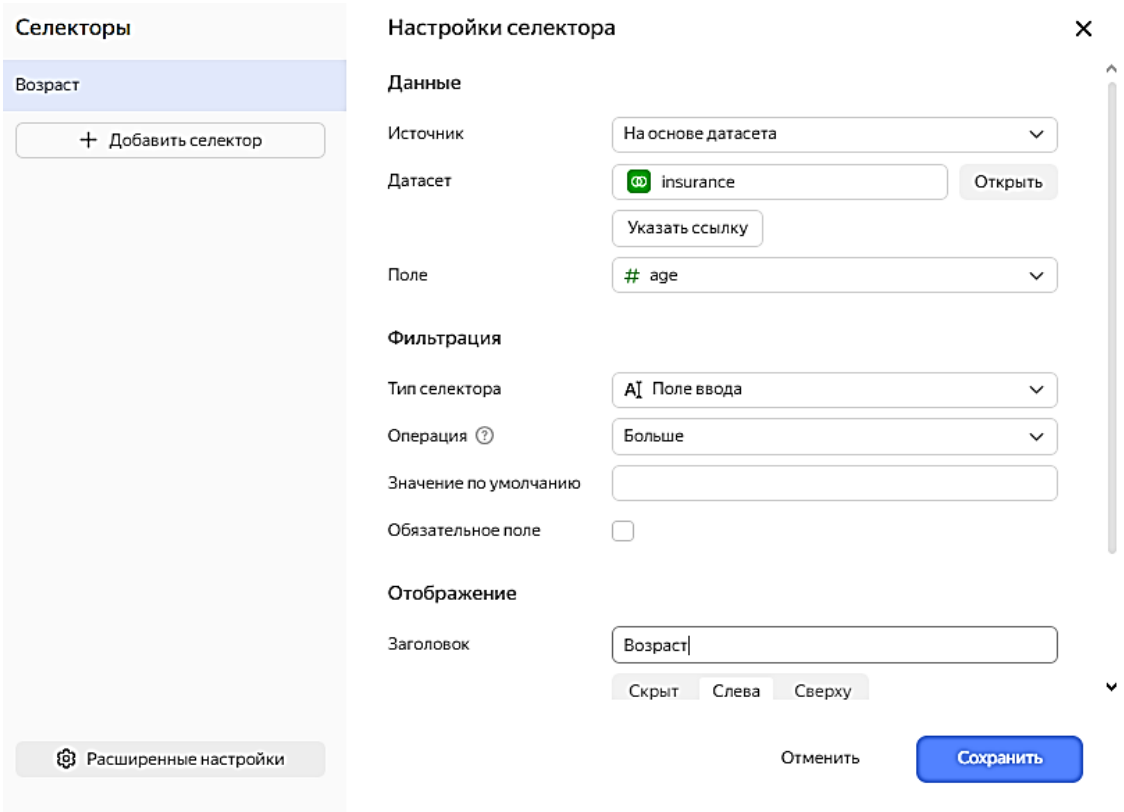


Рис. 2.32. Настройки для селектора по возрасту

Сохраним дашборд (рис. 2.33).

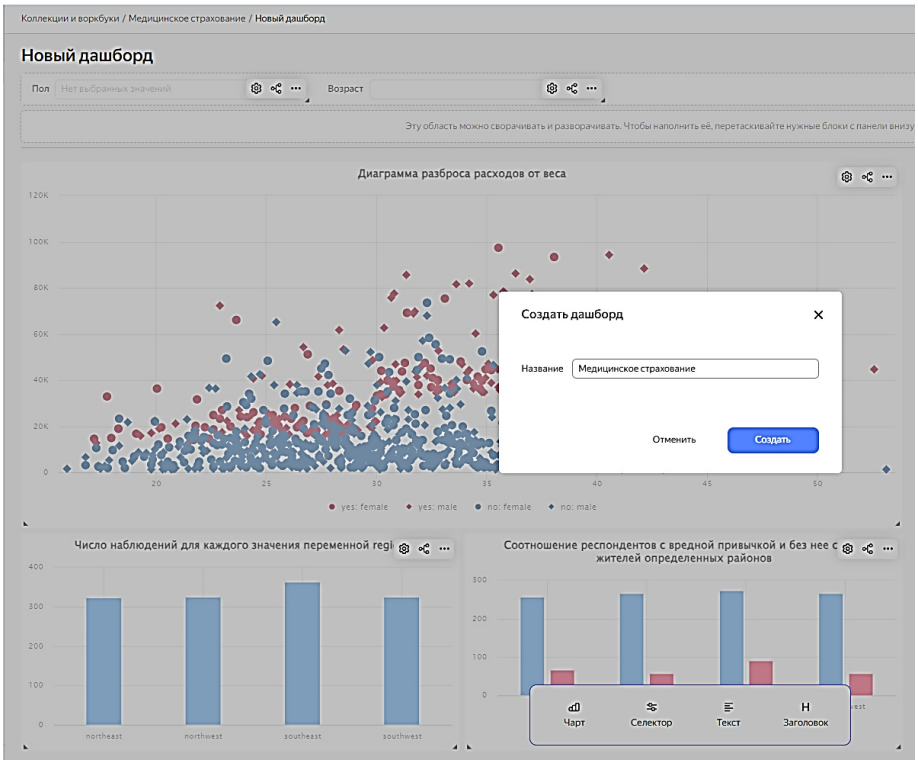


Рис. 2.33. Сохранение дашборда

После сохранения дашборд откроется в режиме просмотра. Применим фильтры и обнаружим, что внесение числового значения в фильтр Возраст – неинформативно, нужно добавить в заголовок и саму операцию сравнения (рис. 2.34).

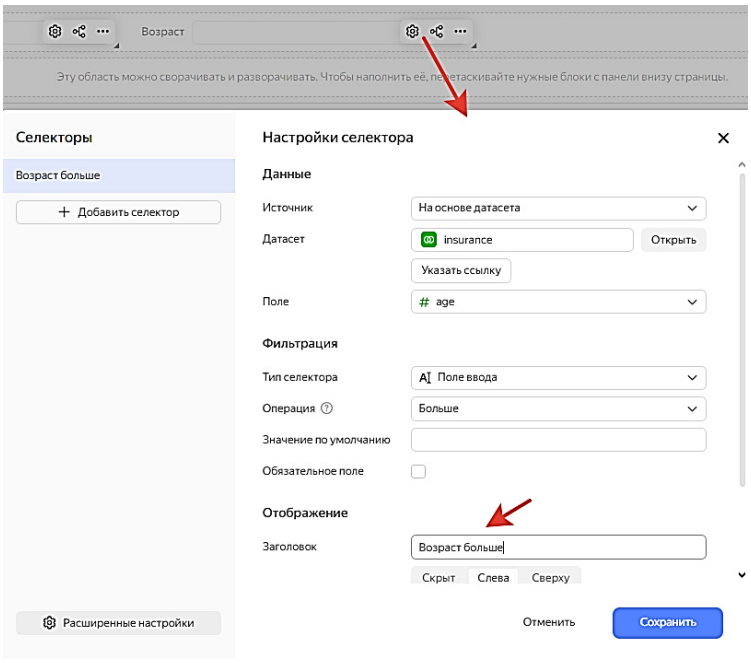


Рис. 2.34. Изменение настроек селектора

На созданный дашборд добавим ссылку (рис. 2.35) на источник данных на ресурсе (https://www.kaggle.com/mirichoi0218/insurance).

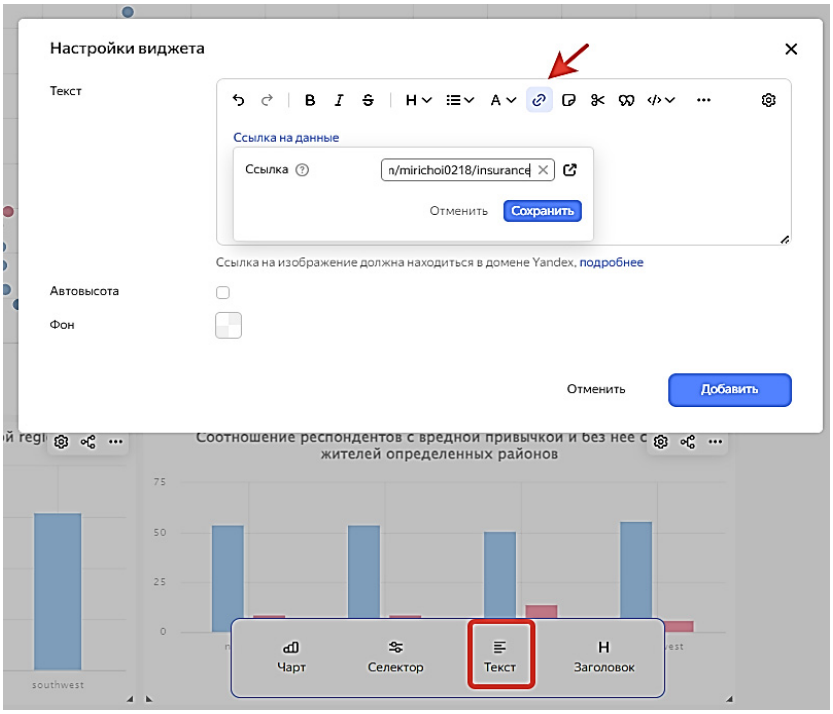


Рис. 2.35. Добавление ссылки на данные

Сохраним дашборд. В режиме просмотра его вид представлен на рис. 2.36.

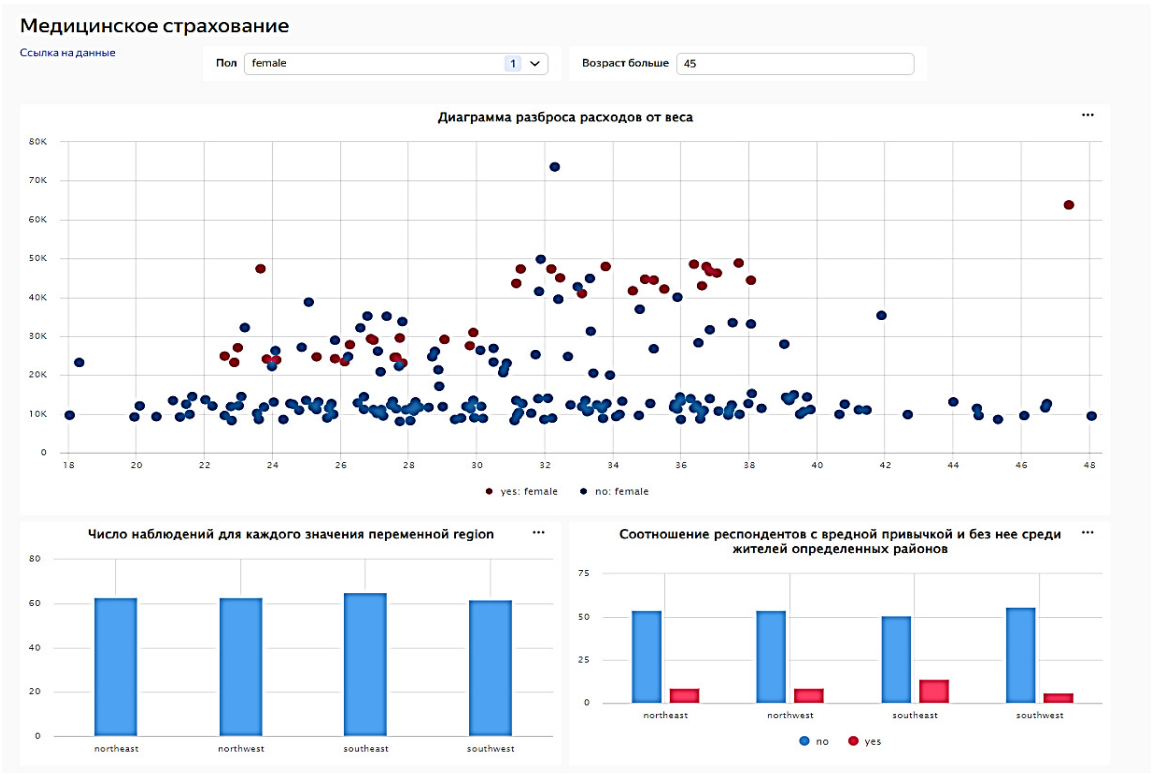


Рис. 2.36. Дашборд в режиме просмотра

Теперь попробуем опубликовать дашборд.

Перед публикацией необходимо убедиться, что:

– чарты и дашборды не содержат персональных данных или коммерческой тайны, так как доступ к опубликованным данным неограничен;

– у вас есть права администратора на публикуемые чарты и дашборды и связанные с ними объекты.

В сохраненном дашборде нажимаем кнопку дополнительных опций рядом с местоположением отчета в иерархии папок (рис. 2.37) и выбираем опцию Публичный доступ.

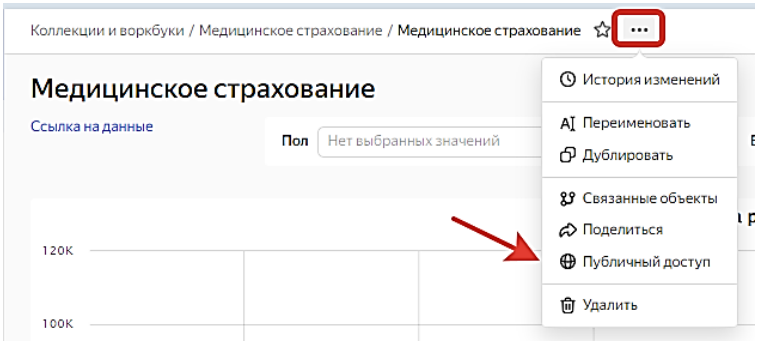


Рис. 2.37. Опция Публичный доступ

Настраиваем публичный доступ (рис. 2.38).

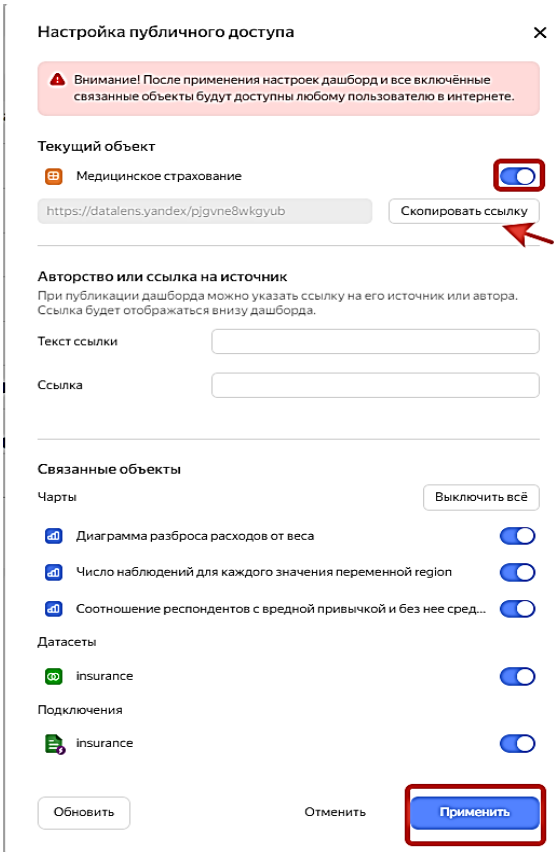


Рис. 2.38. Настройка публичного доступа

Скопировав ссылку и опубликовав ее на каком-либо ресурсе или разослав ее определенному кругу лиц, вы облегчаете доступ к созданному дашборду.

2.5. Объединение данных

Создадим новый воркбук с именем Продажи (рис. 2.39), в котором будем размещать результаты работы с новыми наборами данных.

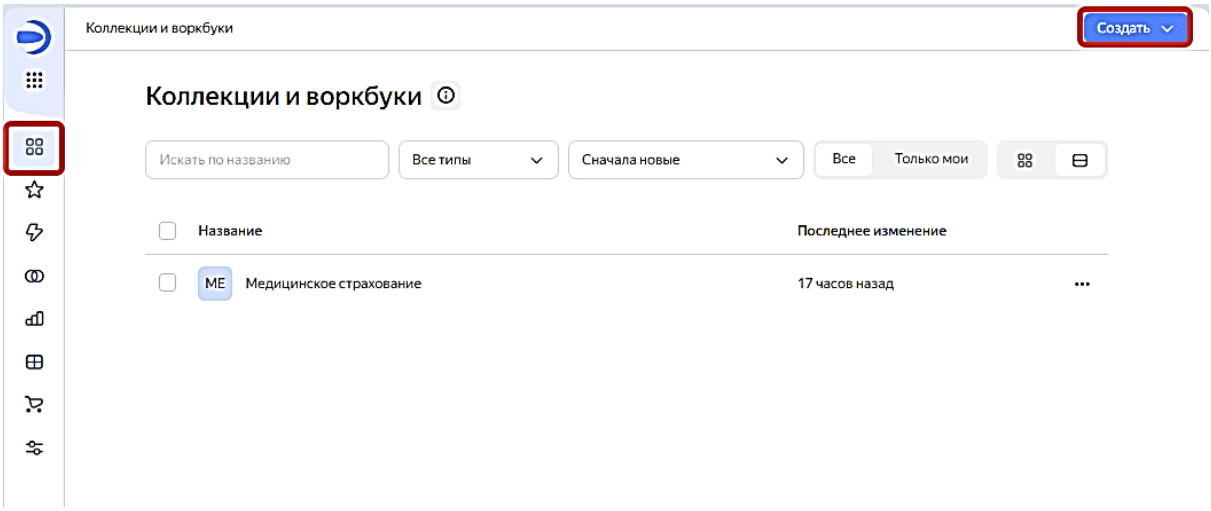


Рис. 2.39. Создание воркбука

Создадим новое подключение, загрузим файл retail.xlsx (рис. 2.40), описание появления которого приведено ранее (см. разд. 1.5).

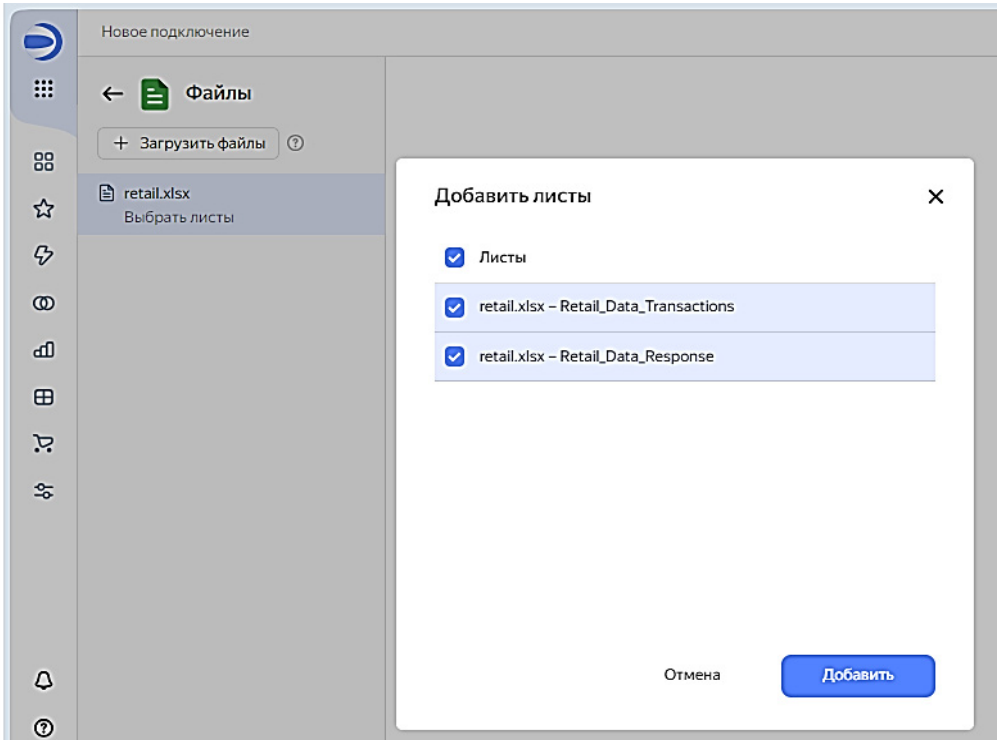


Рис. 2.40. Новое подключение

Сохраним это подключение с именем retail и создадим на его основе датасет (рис. 2.41).

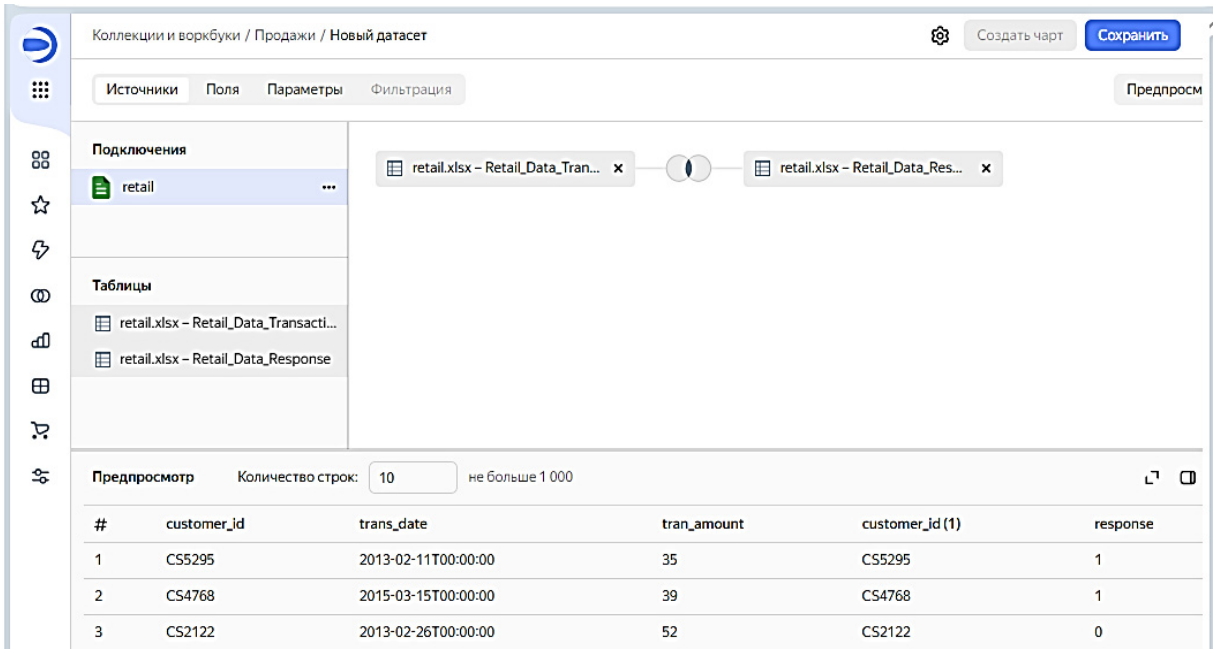


Рис. 2.41. Создание датасета

Так как в этих таблицах есть столбцы с данными об одних и тех же покупателях (Customer_ID), то произошло внутреннее соединение двух таблиц по этому полю (рис. 2.42). К настройкам получаем доступ по щелчку на значке типа соединения.

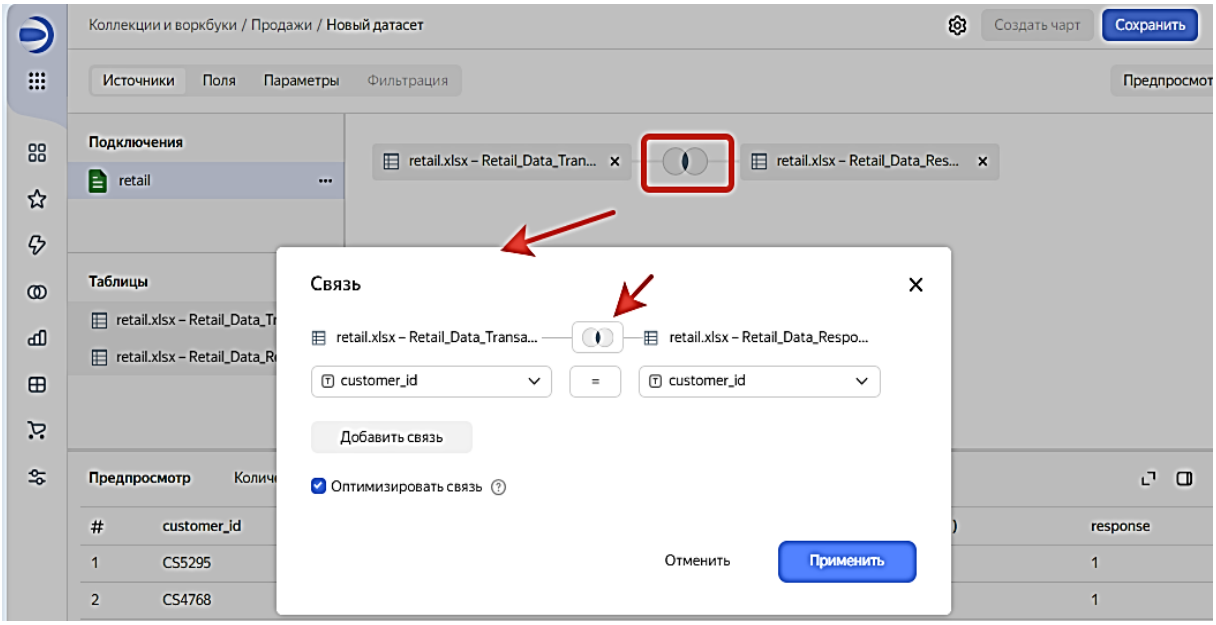


Рис. 2.42. Настройки для связи между таблицами

Перейдем на вкладку датасета Поля и скроем дублирующее поле из второй таблицы (рис. 2.43).

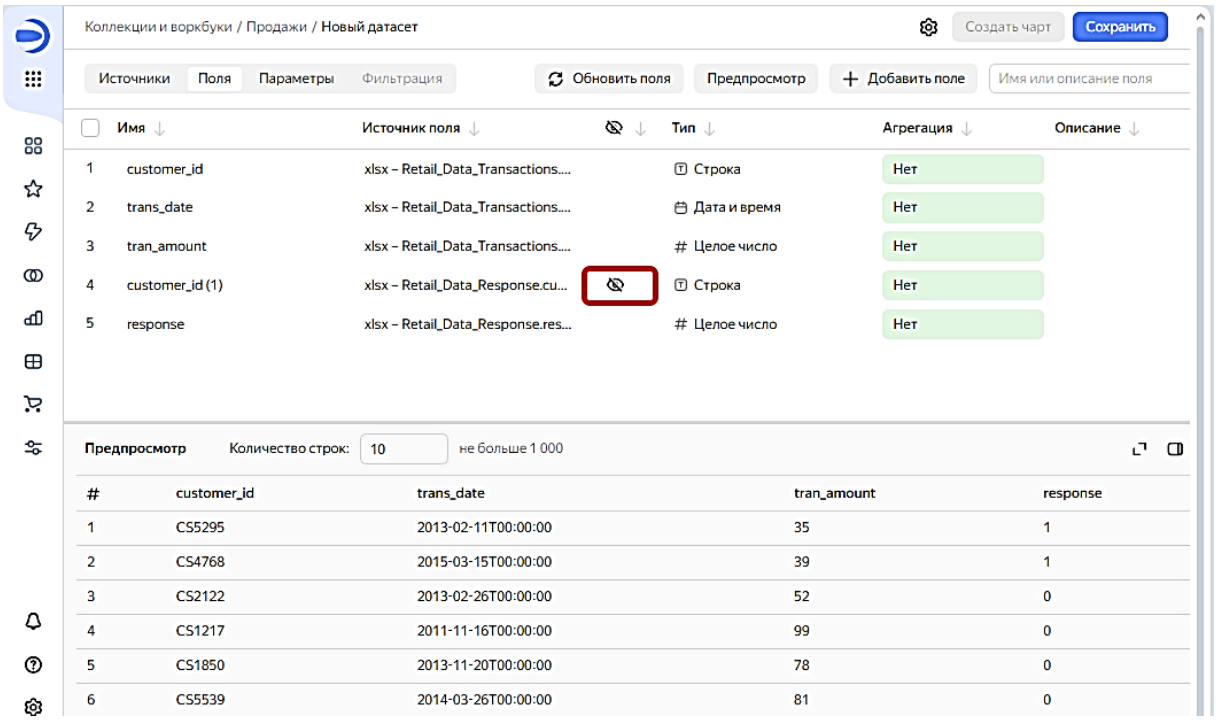


Рис. 2.43. Настройки для полей

Сохраним датасет с именем retail и приступим к созданию чарта (рис. 2.44).

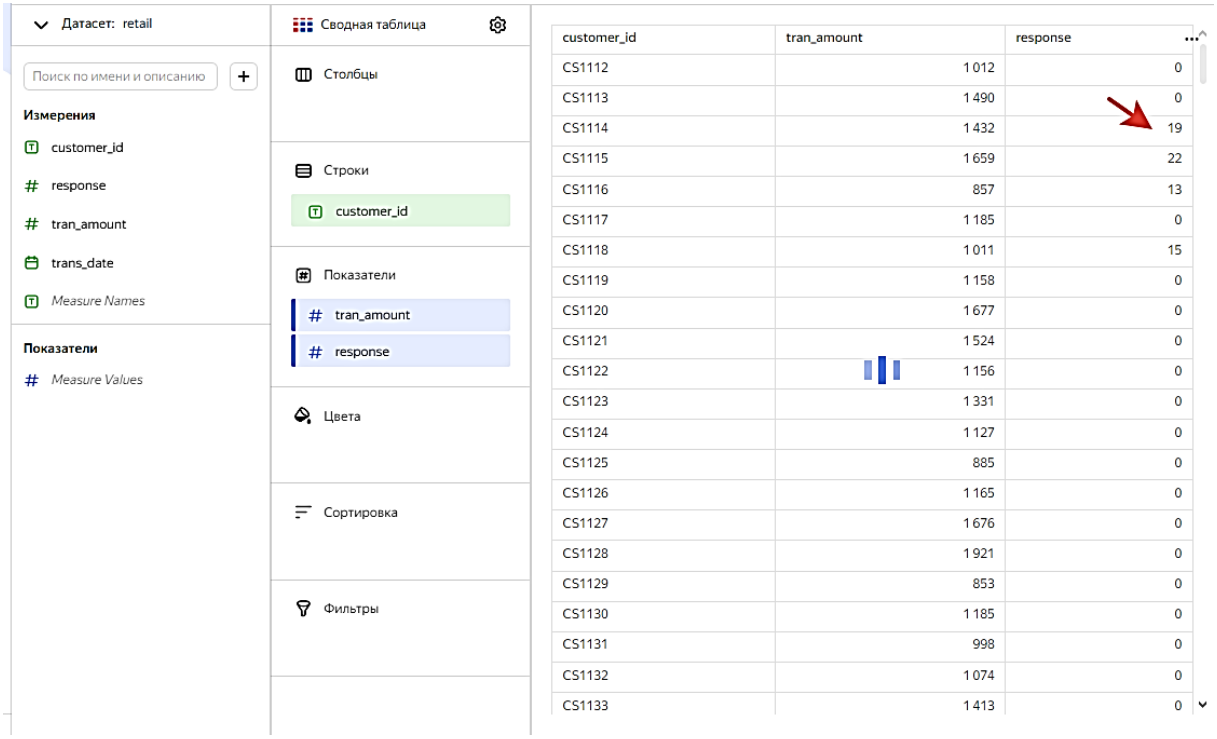


Рис. 2.44. Чарт Сводная таблица

Показатель tran-amount суммирует все покупки данного клиента, но и показатель response также суммирует все 1 или 0, которые относятся к конкретному клиенту. В случае суммирования 0 – все корректно, но значение 19 означает, что клиент откликнулся на рекламу и совершил 19 транзакций. В этом столбце значения по-прежнему должны быть 0 и 1, поэтому кликнем на значке # рядом с именем показателя и в качестве агрегации выберем минимум (можно выбрать и максимум, и среднее, нужный результат будет достигнут), применим настройки (рис. 2.45) и проконтролируем полученный результат (рис. 2.46).

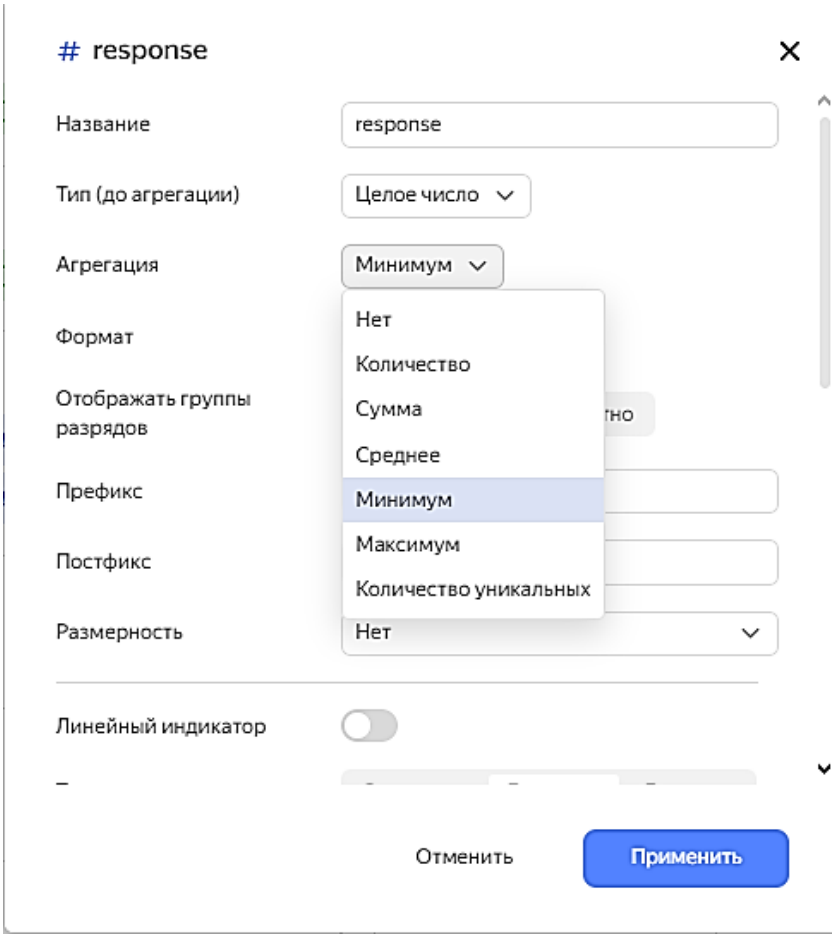
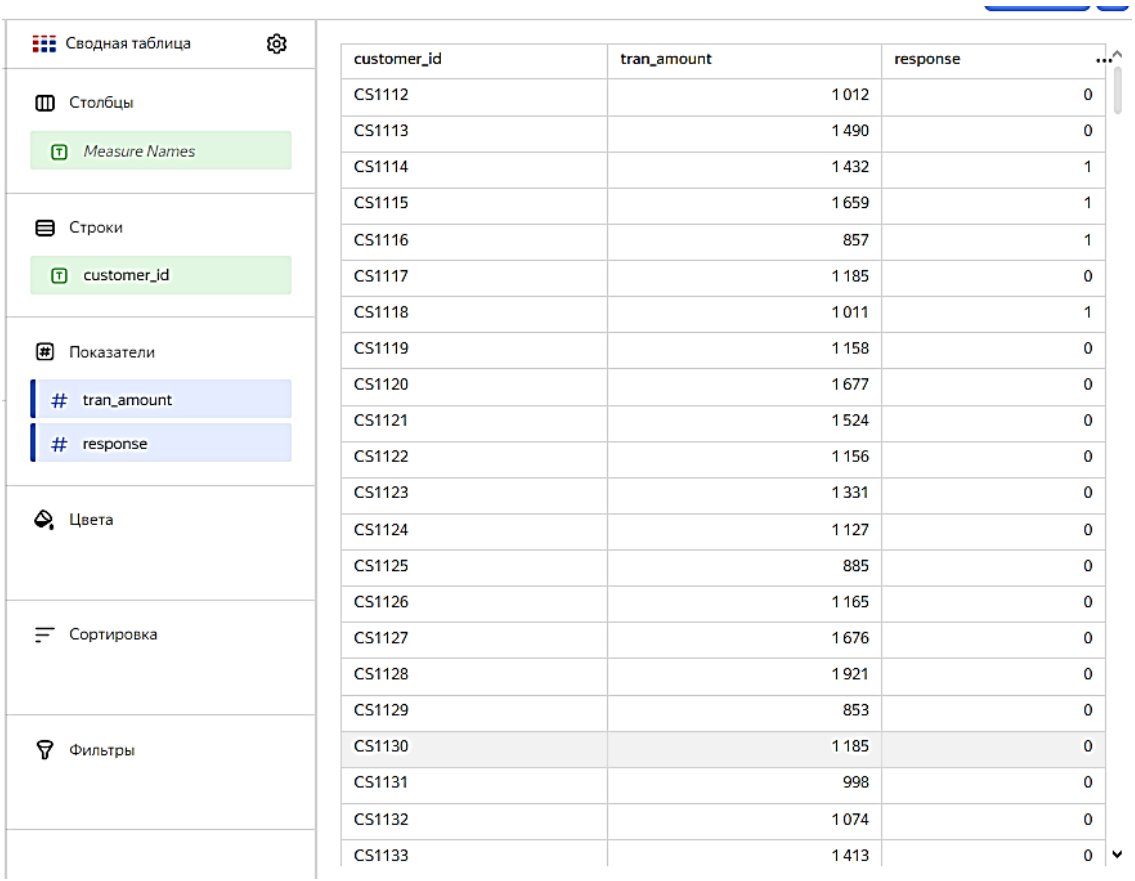


Рис. 2.45. Настройки для показателя response

Ни фильтрация, ни сортировка по агрегированному показателю tran_amount в сводной таблице невозможна. Для выявления наиболее активных клиентов в дальнейшем построим новый датасет, использующий тоже самое подключение retail.

Сохраним чарт и приступим к созданию нового. В новом чарте ось X – trans_date, ось Y – tran_amount. Для показателя trans_date настроим агрегацию по месяцам (рис. 2.47)



customer_id	tran_amount	response
CS1112	1 012	0
CS1113	1 490	0
CS1114	1 432	1
CS1115	1 659	1
CS1116	857	1
CS1117	1 185	0
CS1118	1 011	1
CS1119	1 158	0
CS1120	1 677	0
CS1121	1 524	0
CS1122	1 156	0
CS1123	1 331	0
CS1124	1 127	0
CS1125	885	0
CS1126	1 165	0
CS1127	1 676	0
CS1128	1 921	0
CS1129	853	0
CS1130	1 185	0
CS1131	998	0
CS1132	1 074	0
CS1133	1 413	0

Рис. 2.46. Результат применения настроек

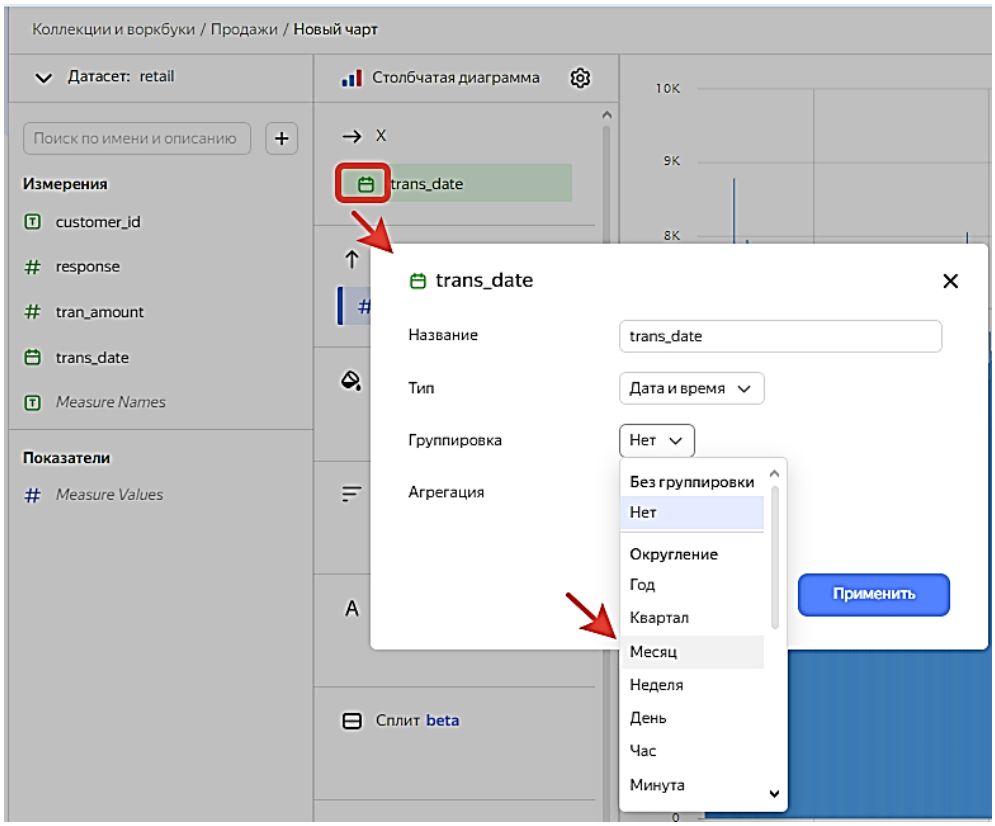


Рис. 2.47. Изменение агрегации для показателя trans_date

Изменим тип диаграммы на линейную и сохраним построенный чарт (рис. 2.48).

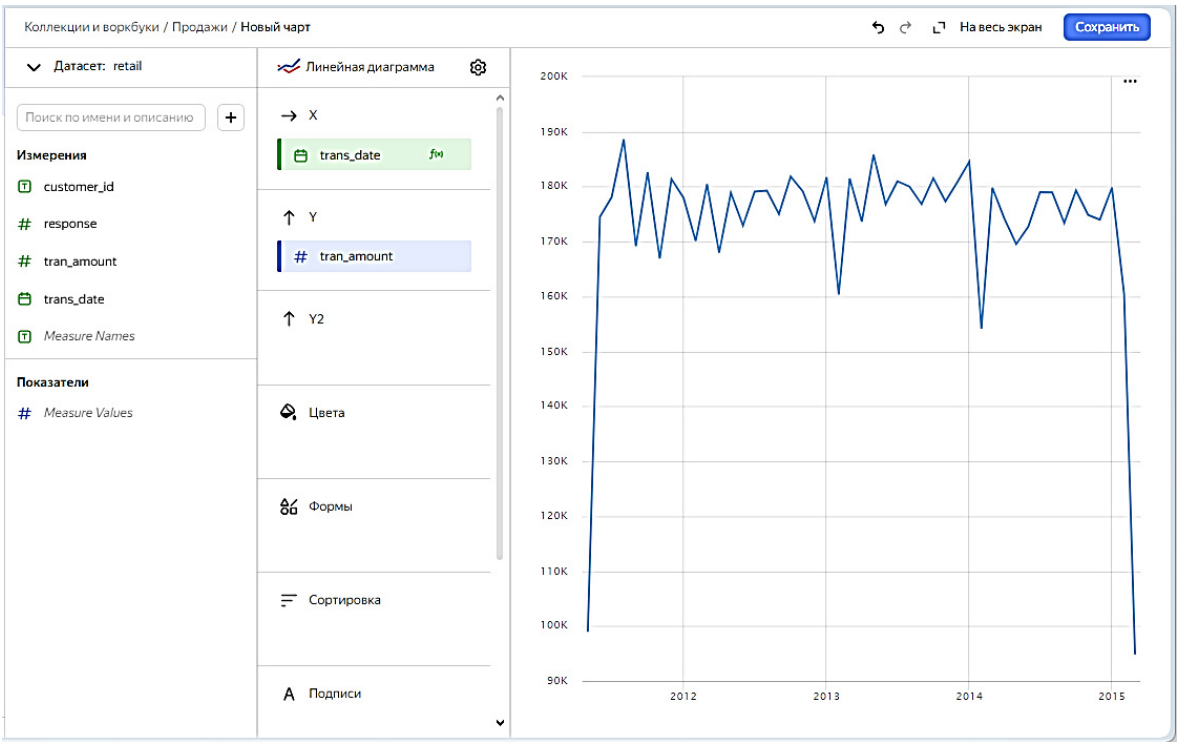


Рис. 2.48. Чарт Линейная диаграмма

Создадим новый дашборд, разместим его в воркбуке Продажи. На дашборд добавим два созданных чарта и настроим селектор по показателю trans_date (рис. 2.49).

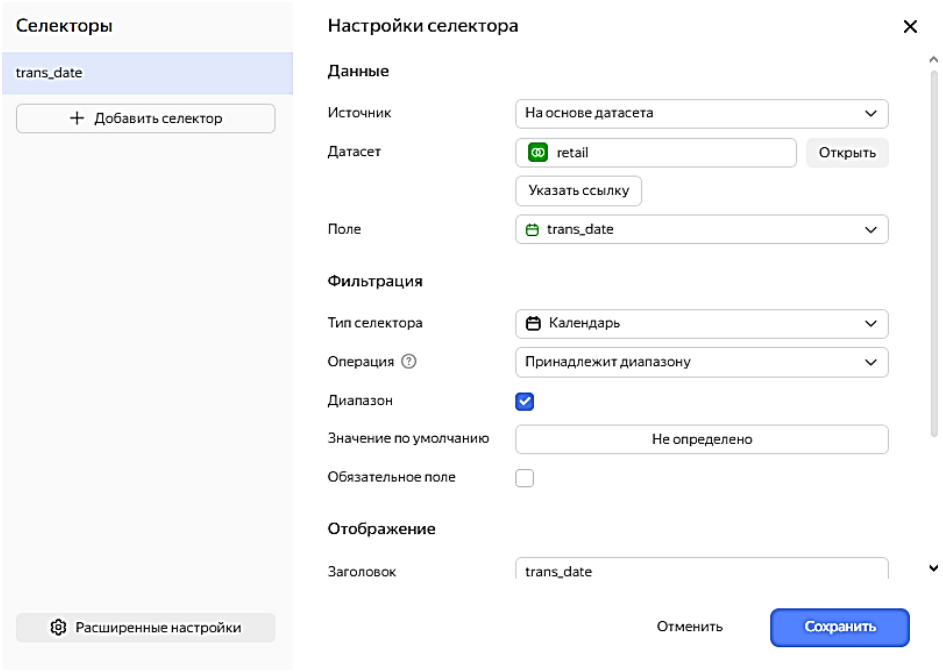


Рис. 2.49. Настройки для селектора

Сохраним дашборд с именем Продажи и настроим диапазон дат для селектора (рис. 2.50).

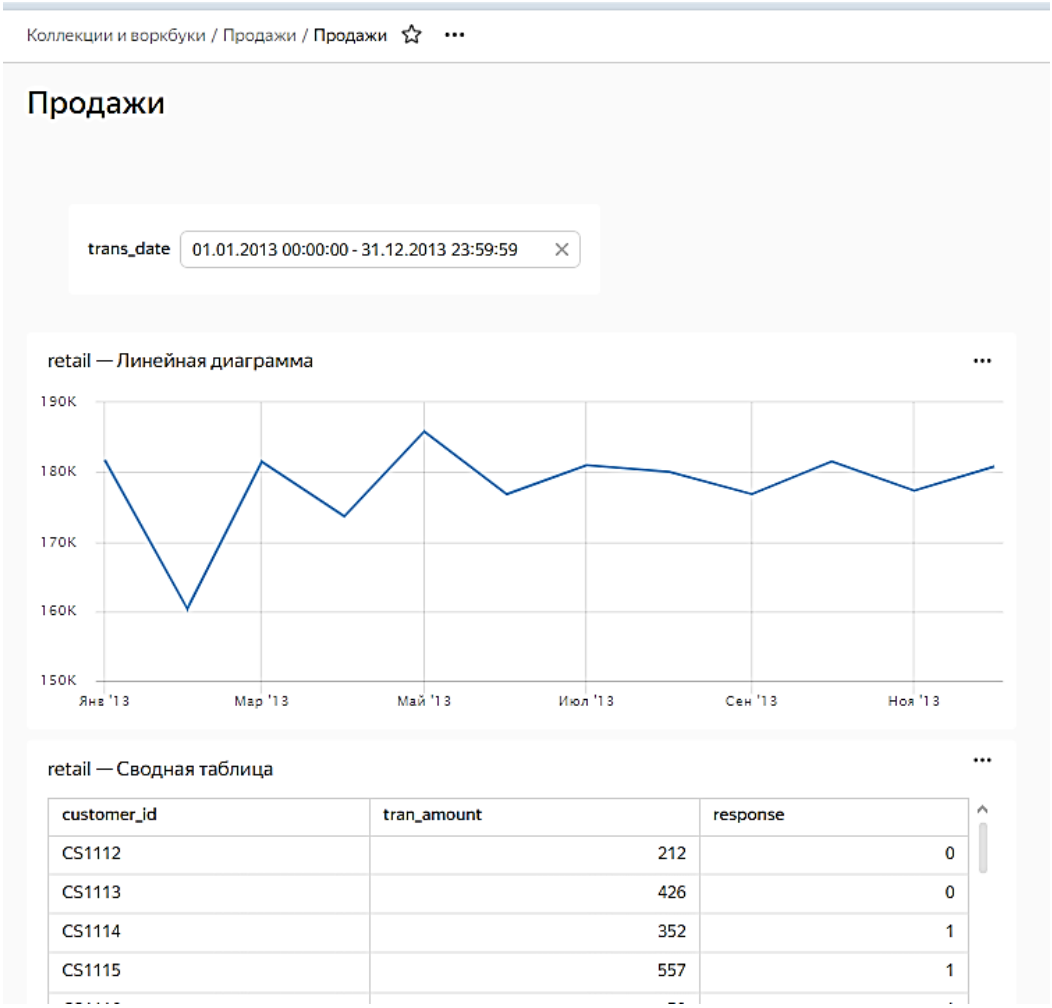


Рис. 2.50. Дашборд Продажи

Создадим новый датасет, открыв подключение retail. В этом датасете также объединим данные и перейдем на вкладку Поля. Скроем два поля trans_date и дублирующее поле customer_id(1), а также настроим агрегацию для полей tran_amount и response (рис. 2.51). Переименуем поле tran_amount (рис. 2.52).

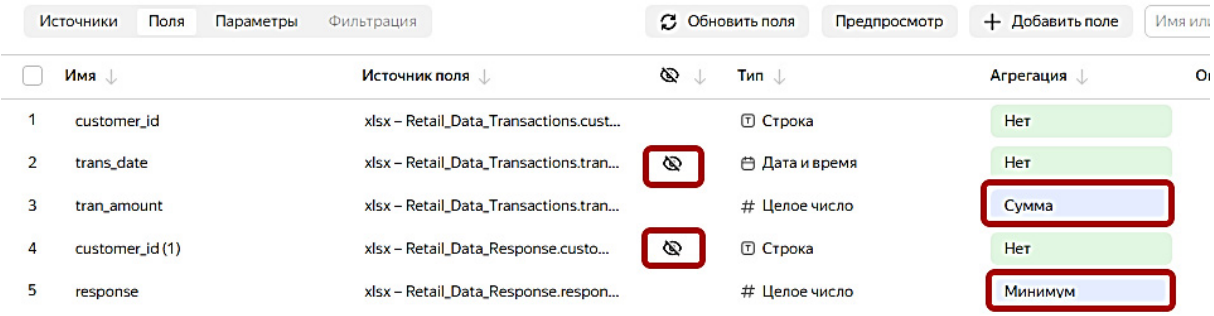


Рис. 2.51. Настройки для полей датасета

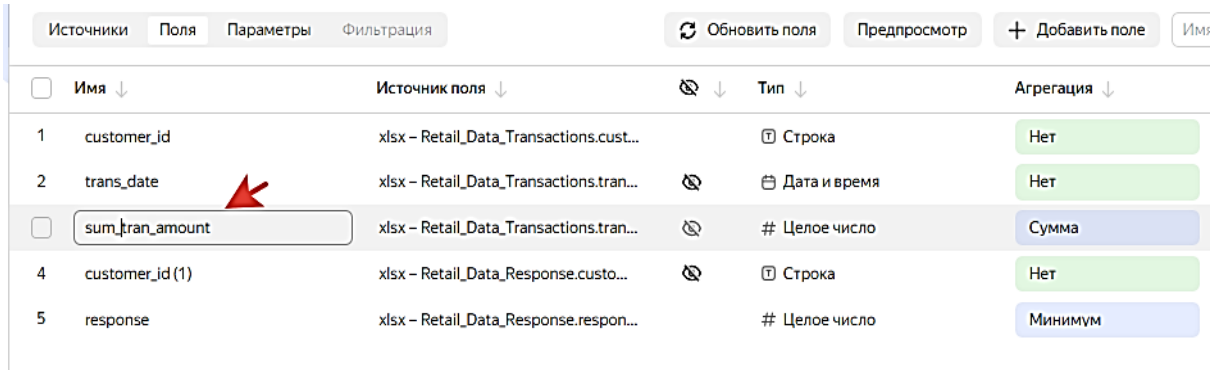


Рис. 2.52. Изменение имени поля

Сохраним датасет с именем retail_ag и создадим на его основе чарт. Тип диаграммы – столбчатая, ось X – customer_id, ось Y – sum_tran_amount, также добавим фильтр для trans_date, оставив в рассмотрении только клиентов с суммарными покупками больше чем 2400 (рис. 2.53).

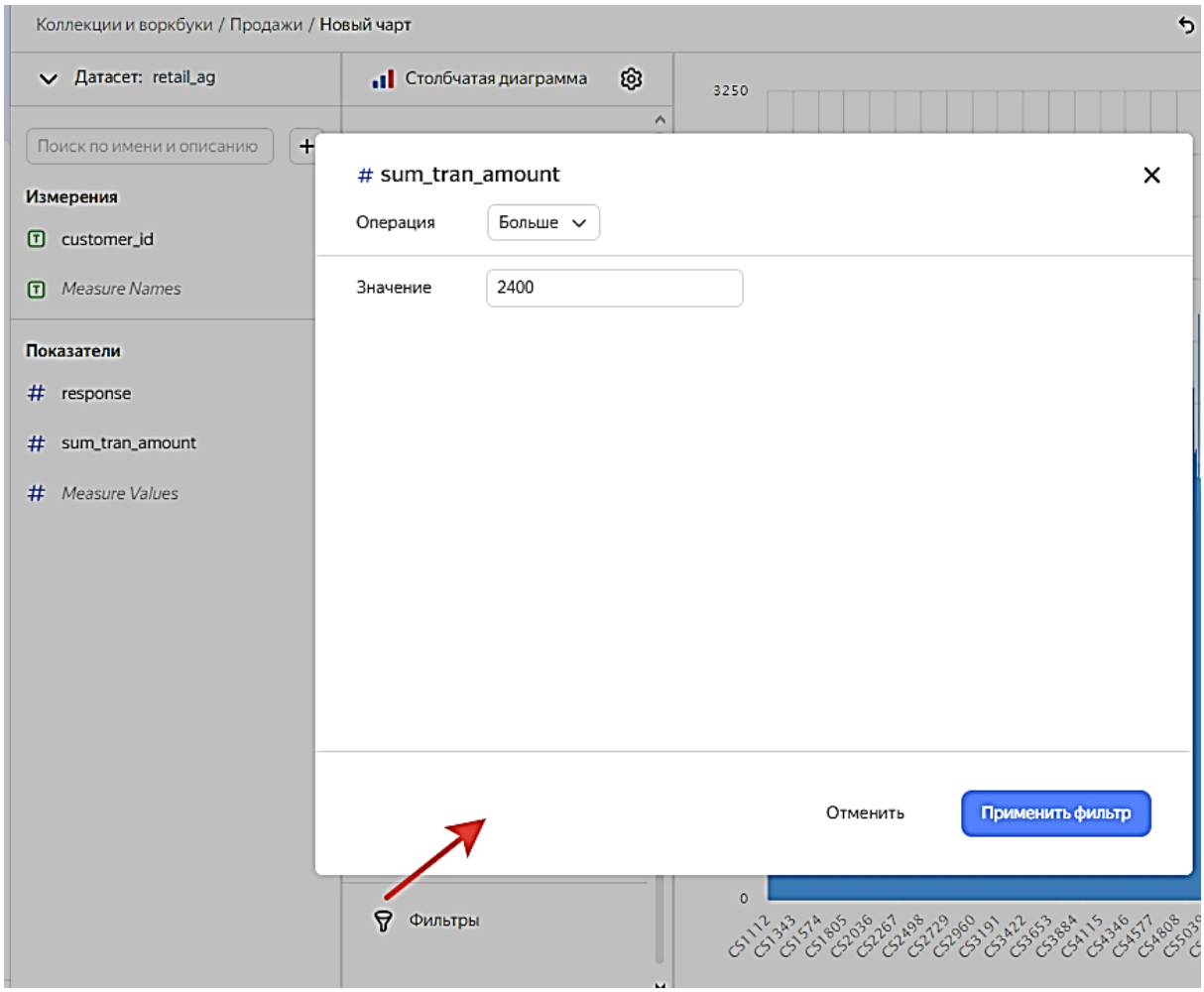


Рис. 2.53. Настройки фильтра

Также добавим настройки для цвета и сортировки (рис. 2.54). Сохраним построенный чарт.

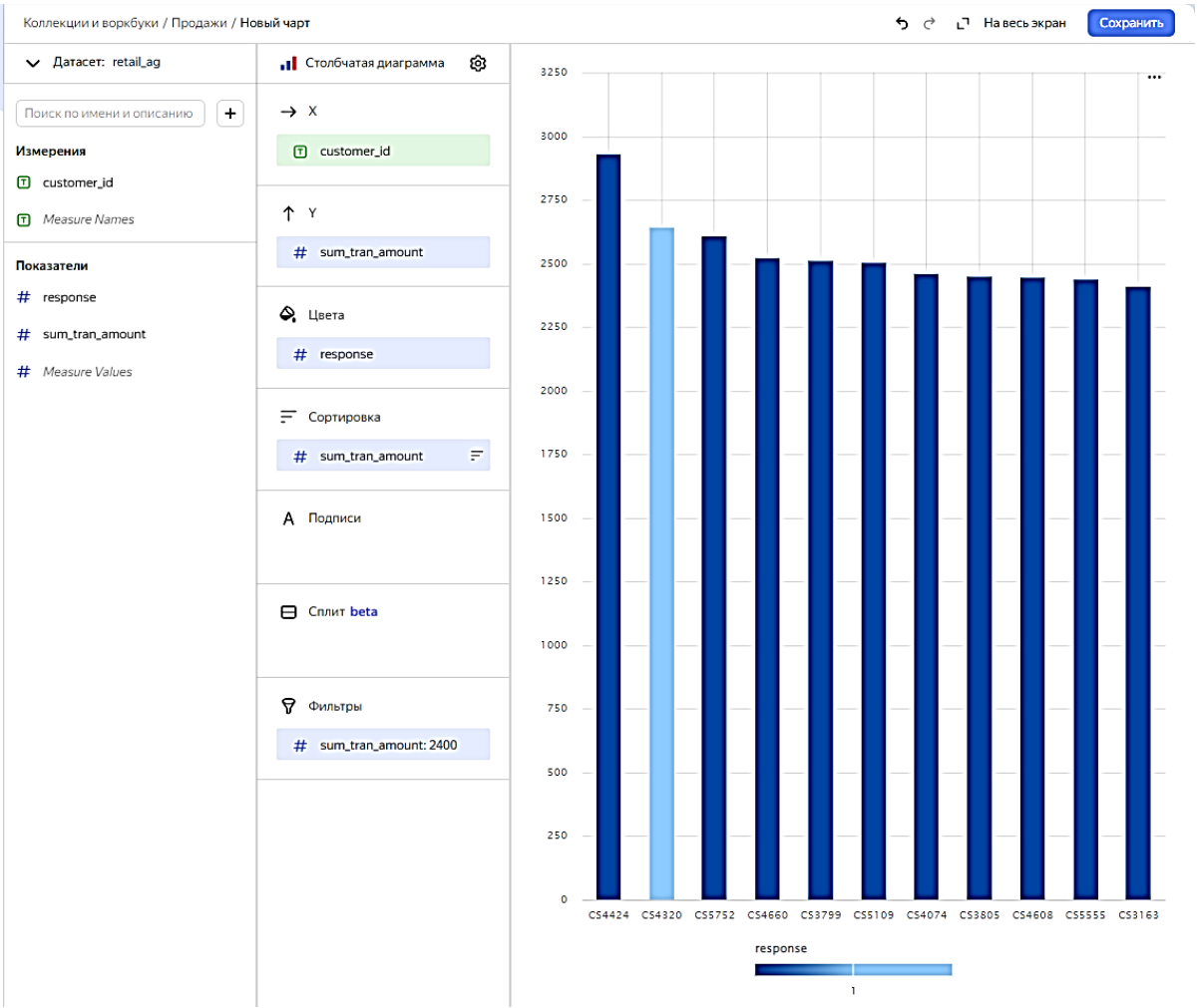


Рис. 2.54. Настройки для чарта

Для дальнейшего совершенствования навыков работы на платформе Yandex DataLens можно воспользоваться справочной документацией [3].

Лабораторная работа № 1

СОЗДАНИЕ ОТЧЕТА ПО РАЗВЕДОЧНОМУ АНАЛИЗУ ДАННЫХ СРЕДСТВАМИ GLARUS BI

Цель работы: для набора данных создать отчет, используя возможности Glarus BI.

Формируемые знания, умения и навыки: научиться загружать данные, работать с вычислимыми полями. Получить навыки формирования отчета по данным, включающего несколько визуализаций, с пользовательскими настройками форматирования.

Необходимо:

1. Загрузить набор данных.
2. Создать запросы: различные виды диаграмм, сводную таблицу. Для сводной таблицы использовать условное форматирование. В одной из диаграмм использовать сортировку.
3. Создать дашборд, на который добавить заголовки и поясняющий текст. Разместить на дашборде построенные запросы, при необходимости скорректировать визуализации (цвет фона, заголовки для визуализаций), добавить на дашборд фильтры.

Контрольные вопросы и задания

1. Как загрузить данные в Glarus BI? Как просмотреть корректность определения типов данных для полей?
2. Что позволяет сделать опция сканирования для загруженного набора данных?
3. Какие визуализации включены в дашборд? Какую бизнес-проблему описывает дашборд? Какие элементы форматирования использовались для повышения наглядности представления данных дашборда?
4. Как создать сводную таблицу?
5. Как создать фильтр?
6. Как скачать дашборд в виде файла с расширением .pdf?

Лабораторная работа № 2

СОЗДАНИЕ ОТЧЕТА ПО РАЗВЕДОЧНОМУ АНАЛИЗУ ДАННЫХ СРЕДСТВАМИ YANDEX DATALENS

Цель работы: для набора данных создать отчет, используя возможности BI-системы от компании Yandex.

Формируемые знания, умения и навыки: научиться подключать источники данных, работать с вычислимыми полями. Получить навыки формирования отчета по данным, включающего несколько визуализаций, с пользовательскими настройками форматирования.

Необходимо:

1. Подключить источники данных.
2. Создать дашборд, использовать следующие чарты: сводная таблица, индикатор и диаграммы. Добавить селекторы. Использовать несколько вкладок в дашборде, также добавить заголовки и поясняющий текст.
3. Опубликовать дашборд.

Контрольные вопросы и задания

1. Как выполнить подключение?
2. Как настроить различные агрегации для показателей и дат?
3. Какие чарты включены в дашборд? Какую бизнес-проблему описывает дашборд? Какие элементы форматирования использовались для повышения наглядности представления данных дашборда?
4. Как создать сводную таблицу?
5. Как создать селектор?
6. Как опубликовать созданный дашборд?

СПИСОК ЛИТЕРАТУРЫ

1. Glarus BI : офиц. сайт. URL: https://glarus-bi.ru/ (дата обращения: 10.01.2025).

2. Документация Glarus BI : офиц. сайт. URL: https://glarus-bi.ru/documentation/index.html (дата обращения: 10.01.2025).

3. Yandex DataLens : офиц. сайт. URL: https://datalens.yandex.ru/ (дата обращения: 10.01.2025).

Учебное издание

Рындина Светлана Валентиновна

Анализ данных: построение отчетов с Yandex DataLens и Glarus BI

Редактор *А. Г. Темникова*
Технический редактор *Н. В. Иванова*
Компьютерная верстка *Н. В. Ивановой*

Подписано в печать 05.05.2025.
Формат 60×84 $^{1}/_{16}$. Усл. печ. л. 5,12.
Тираж 6. Заказ № 146.

Издательство ПГУ
440026, г. Пенза, ул. Красная, 40.
Тел.: (8412) 66-60-49, 66-67-77; e-mail: iic@pnzgu.ru

ПензГУ