# Лабораторная работа «Визуализация и анализ данных с помощью

**Yandex DataLens»**

*Основная цель:*

Научиться использовать Yandex DataLens.

Изучите основные возможности отечественных программных средств интеллектуального анализа данных (ИАД), описанные ниже.

# Теоретическое введение.

**Visiology, Analytic Workspace, Modus BI, Luxms BI, «Форсайт», «Триафлай», PolyAnalyst и Yandex DataLens** — самые популярные российские BI-системы. Они предлагают широкий спектр решений на основе лучших практик западных BI-платформ Power BI, Tableau и Metabase. Заметную позицию на рынке занимает также китайская BI- система Fine BI, которая, впрочем, не входит в реестр российского ПО и не может в полной мере решать задачи импортозамещения. Это позволяет подобрать подходящую функциональность с учетом уникальных BI-задач и размеров бизнеса в разных технологических и ценовых категориях.

Хотя некоторые могут относиться к российским BI-продуктам скептически, этот сегмент активно развивается уже несколько лет. Так, аналитики допускают, что в 2023 году рост рынка может составить 40% по сравнению с 2022, что говорит об увеличивающемся спросе и количестве внедрений BI-решений российских вендоров, а также усовершенствовании их функциональности.

## Обзор и сравнение функциональности российских BI-решений

Чтобы провести комплексный анализ платформ аналитики, мы по нескольким параметрам сравнили популярные продукты BI, такие как Visiology, Yandex DataLens, Analytic Workspace и другие.

1. **Архитектура**

Внедрение BI-системы становится необходимо бизнесу, когда объемы постоянно пополняющихся данных, в том числе неструктурированных, возрастают настолько, что управление ими в рамках условных Excel-таблиц и ручного анализа становится не только затруднительным, но и приводит к ошибкам и неверной интерпретации. В свою очередь, хранение больших данных и работа с ними — это серьезный технологический вызов, связанный в том числе с мощностью используемых серверов.

На данном этапе бизнес, как правило, имеет свою сложную ИТ-инфраструктуру (обычно изобилующую «костылями» и «заплатками»). Внедрение новой «надстройки» в виде BI- системы призвано упростить и структурировать работу с данными, а не разрушить окончательно. С этой точки зрения архитектура современного BI-решения должна быть такой, чтобы более-менее бесшовно встраиваться в существующую инфраструктуру Заказчика, в то же время позволяя закрыть все задачи по бизнес-аналитике данных:

* + сбор и первичная обработка данных;
  + извлечение, преобразование и загрузка данных;
  + их складирование и представление в витринах данных;
  + анализ и его отображение на web-портале.

Чаще всего современные BI-решения построены на микросервисной архитектуре, которая

позволяет использовать мощности нескольких небольших серверов и обеспечить их

«синхронную» работу для решения одной бизнес-задачи.

## Что предлагают BI-решения ?

**Visiology:** платформа построена на микросервисной архитектуре и развертывается в виде набора Docker-контейнеров. Frontend архитектура реализована с помощью JavaScript (с подключением любых соответствующих библиотек), JQuery, Angular, Highcharts, D3.js и собственного документированного JavaScript API.

**Modus BI:** Back-End — язык разработки Golang, Front-End — кросс браузерный ReactJS с адаптивной версткой. Система имеет два независимых модуля — Modus ETL для решения сложных задач обработки данных из различных источников и Modus BI для формирования аналитических отчетов. При этом последний может осуществлять операции только с данными, находящимися в одном источнике.

**Analytic Workspace:** предлагает микросервисную архитектуру из open-source компонентов, которые гибко настраиваются и встраиваются в инфраструктуру Заказчика. Более того, в ней реализованы все слои работы с данными, включая полноценный ETL- модуль, что покрывает весь процесс от сбора и обогащения до визуализации данных.

**PolyAnalyst:** позволяет добавлять пользовательские процедуры для загрузки, преобразования и выгрузки данных на языках R и Python. Также предоставляется возможность обрабатывать и анализировать данные, применяя алгоритмы машинного обучения.

## Подключение к источникам данных

Качественная BI-система должна уметь подключаться к очень большому количеству источников: разнообразные базы данных, файлы различных форматов (как локальные, так и удаленные, доступные по ftp-подобным протоколам), web-сервисы. Сравним возможности российских BI-решений в части подключения к популярным источникам данных.

На рисунке ниже представлена информация о доступных для подключения источниках данных в рассматриваемых BI-решениях (Рисунок 1)

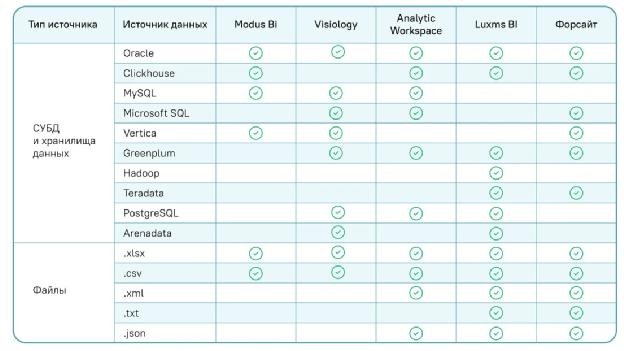


Рисунок 1 — Источники данных BI-решений

## Трансформация и хранение данных (наличие ETL-модуля)

Чтобы эффективно работать с данными из разных источников применяется набор механизмов, которые обычно называются модулем ETL (Extract-Transform-Load). Стадия Extract обеспечивает выгрузку данных. На стадии Transform происходит их объединение в единую модель данных, очистка и обогащение. Стадия Load выполняет загрузку в постоянное производительное хранилище аналитических данных.

Если же BI-решение не оснащено инструментами обработки данных, приходится подключать сторонние. Это усложняет и тормозит работу с большим массивом данных, особенно если он постоянно обновляется.

Часто ETL-слой не закрывает все задачи и достаточно ограничен, особенно при работе с большим объемом данных:

**«Триафлай»:** нет возможности присоединять таблицы из разных источников и дополнительно обрабатывать их на этапе трансформации.

**Modus BI:** также нужно формировать отдельный датасет в виде таблицы для каждой диаграммы. При этом потребуется знание языка SQL.

**Visiology:** встроен базовый ETL, ориентированный на загрузку по расписанию подготовленных витрин данных. Тем не менее система советует подключить внешний ETL-инструмент для более сложных трансформаций.

**Yandex DataLens:** полностью отсутствует ETL-слой. Задачи можно решить с помощью других инструментов — Yandex Data Transfer, Yandex DataProc, Yandex DataSphere и т. Д.

**Analytic Workspace:** реализован **полноценный ETL-слой**, который включает

подсистемы источников данных, моделей, виджетов, информационных панелей и администрирования. Также можно предварительно просмотреть результаты моделирования, что ускоряет работу с большими объемами данных.

## Гибкость системы при работе с данными и расширенная аналитика

Этот критерий оценивает функциональные возможности работы системы с подготовленными данными и использования их для решения аналитических задач. Мы рассмотрим принципы и ограничения работы и с моделью данных и при создании аналитических показателей на их основе в различных BI-продуктах.

## «Форсайт»:

имеет возможности расширенной аналитики данных: более 300 встроенных алгоритмов продвинутой аналитики. Среди них базовые математические и статистические методы, анализ временных рядов, корреляционно-регрессионный анализ, Text Mining, Data Mining, многомерный статистический анализ, распознавание естественной речи и графических образов, имитационное моделирование и экспертные методы.

**Analytic Workspace:** работает на базе ClickHouse + Apache Spark, позволяет агрегировать большие объемы данных, обрабатывать SQL-запросы со множеством готовых операторов и масштабировать систему с помощью drag-n-drop, а также за счет подключения разнородных источников данных.

**Luxms BI:** имеет значительные возможности для обработки и анализа данных вплоть до кастомной визуализации по эскизам Заказчика. Структура работы с данными предполагает возможность как выполнения расчетов "на лету", так и предрасчета данных на стороне внутренних/внешних источников. Однако функциональность построения моделей данных и настройки аналитических объектов требует серьезного участия разработчиков.

**Visiology:** есть возможность работать не только в In-Memory, но и в режиме push-down автоматической генерации SQL-запросов к распределенным СУБД c массовым параллелизмом (Arenadata, HP Vertica, Yandex ClickHouse): «горячие» данные размещаются в оперативной памяти, быстро обновляются и доступны для обработки без задержек; «холодные» ретроспективные данные содержатся в реляционном хранилище .

**Modus BI:** в модуле Modus ETL реализованы сценарии стандартных операций по обработке данных и настроены шаблоны, которые позволяют значительно упростить настройку ETL-процесса. Чтобы реализовать более сложные сценарии обработки или загрузки данных, может требоваться написание операций на языке 1С/SQL .

## PolyAnalyst:

ограничены возможности настройки определенных визуальных элементов (например, OLAP-таблиц).

**«Триафлай»:** анализирует версионности и ретроспекцию, показывает, как было получено значение, углубляясь в формулу расчета показателя, наложенные фильтры, разрезы агрегирования и даже исходные данные.

## Удобство использования и простота освоения (Self-service)

«Селф сервис» (self-service) системы — это так называемые BI-системы самообслуживания, которые позволяют найти решение проблемы без службы поддержки. Такие платформы подходят для специалистов без значительных ИТ-знаний, часто легко интегрируются и быстро дают результат.

Хотя self-service частично реализуется у многих платформ, не все можно отнести к low- code/no-code (LCNC) решениям. Самых впечатляющих результатов в этом отношении добились разработчики системы AnalyticWorkspace. Например, Visiology, которая чаще развертывается на ресурсах клиентов (on-premise), потребует больше ресурсов разработки по сравнению с Analytic Workspace, которую можно интуитивно развернуть в облаке и быстро приступить к анализу данных.

**Analytic Workspace**: интуитивно понятный интерфейс, в котором можно загружать данные, создавать модели и массивы, проводить разведку данных, визуализировать их и создавать виджеты и дашборды (информационные панели) с помощью drag-n-drop.

**Visiology:** запустила виртуального помощника ViTalk, который строит дашборды

(информационные панели) по голосовым командам.

## Совместимость с российским ПО

Чтобы обеспечить безопасность работы с данными и высокую производительность BI- систем, они должны быть совместимы с российским ПО. Более того, с развитием тренда на импортозамещение некоторым компаниям приходится выбирать только совместимое с российскими процессорами и системами ПО, чтобы соблюдать требования Минцифры.

В единый реестр российского ПО включены:

* + Visiology;
  + Analytic Workspace;
  + Yandex DataLens;
  + Modus BI;
  + Power BI;
  + Luxms BI;
  + «Форсайт»;
  + «Триафлай»;
  + PolyAnalyst.

Достаточно зрелая китайская платформа Fine BI не внесена.

## Интерфейс платформы и порог входа

Хотя почти все российские системы аналитики предлагают достаточную функциональность по работе с данными, у многих из них высокий порог входа: специалисту без технического опыта будет сложно разобраться и настроить процессы. Чтобы ускорить и упростить внедрение, особенно в компаниях без опытных аналитиков данных, лучше выбирать интуитивные решения с простым интерфейсом и настройкой.

**«Форсайт»:** нужно освоить внутреннюю логику построения отчетов и изучить документацию.

**Visiology:** из-за дополнительной пользовательской настройки требует значительных ресурсов при развертывании.

**Yandex DataLens:** понятный интерфейс для визуализации данных.

**Analytic Workspace:** интуитивно понятный интерфейс, чтобы собирать, обрабатывать и визуализировать данные. Благодаря функции drag-n-drop можно гибко и просто настраивать отчеты.

## Гибкое ценообразование

Запросы бизнеса часто меняются в процессе работы с данными, что заставляет расширять или, наоборот, сужать требования к BI-системе: функциональность, количество пользователей, объем хранилища, типы данных и их объем и другие. Чтобы оптимизировать инвестиции, при этом не потеряв в производительности, стоит выбирать решения с гибкими и доступными ценами.

BI-вендоры предлагают развернуть системы как on-premise (на ресурсах клиента), так и в облаке. Цена BI решения часто рассчитывается по запросу, но можно оформить подписку за пользователя:

**Visiology:** от 35 тыс. руб. за пользователя в год.

**Analytic Workspace:** от 5 тыс. руб. за пользователя в год в облачной версии.

**«Форсайт», Luxms BI:** предоставляют цену только по запросу и за конкретный проект (от 2 000 000 руб)

Бесплатный инструмент для визуализации — **Yandex DataLens**, однако он не закрывает весь цикл работы с данными. Чтобы оценить BI-систему, можно воспользоваться демодоступом, например, в Analytic Workspace он бесплатный в течение 14 дней.

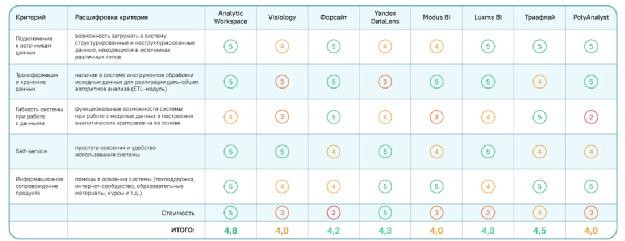
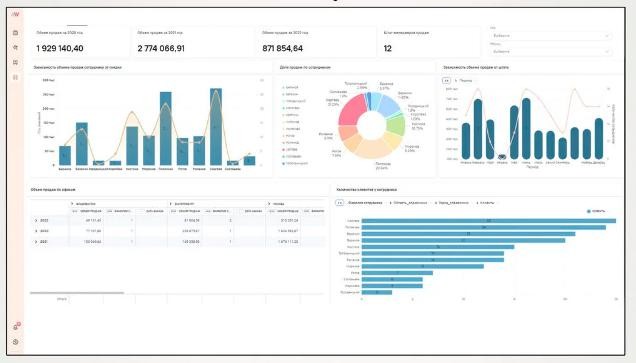
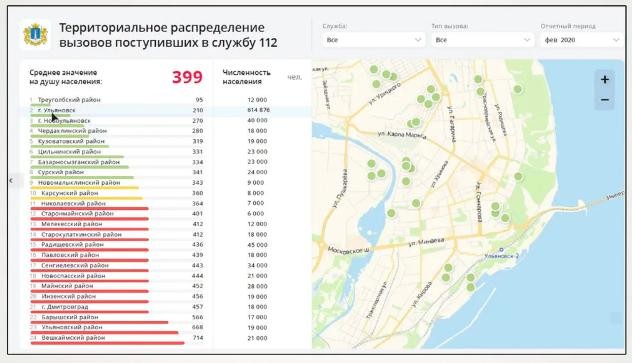
На рисунке ниже приведены параметры BI-решений и их оценки (Рисунок 2)

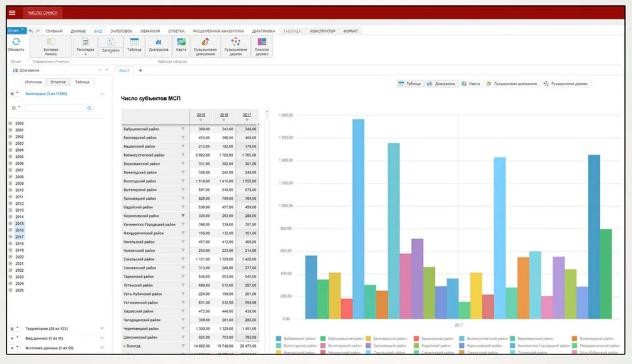
Рисунок 2 — параметры BI-решений и их оценки

**Примеры дашбордов (информационных панелей) российских BI-систем Analytic Workspace**

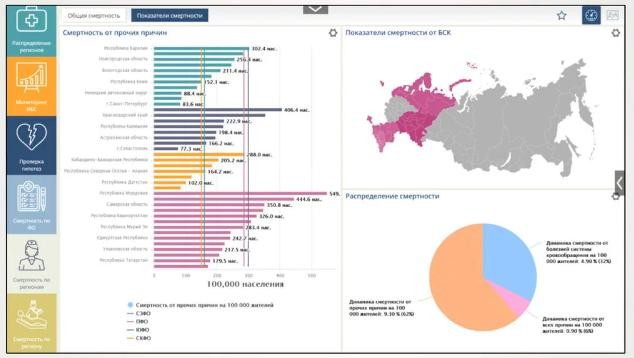


**Visiology**

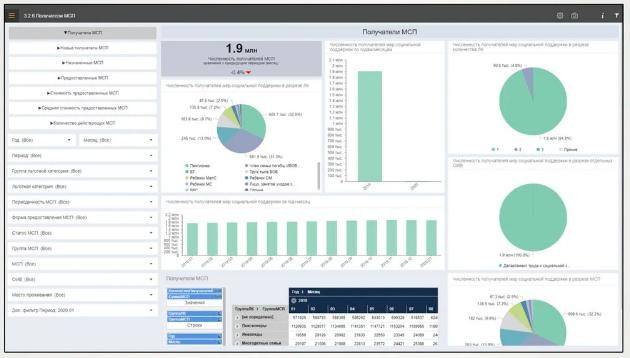


**Форсайт **

**Luxms Bi**



**Modus**



**Выводы**

Чтобы понять, какую систему аналитики выбрать, необходимо определить задачи по работе с данными. Например, если нужна только визуализация данных, Yandex Data Lens

— хороший и доступный вариант во внешнем контуре (Yandex Data Lens не может встраиваться во внутреннюю инфраструктуру Заказчика). Если же нужно закрыть задачи по сбору, хранению, трансформации и визуализации, стоит рассмотреть комплексные платформы.

# Задание.

Чтобы начать работать с DataLens выберите одно из двух:

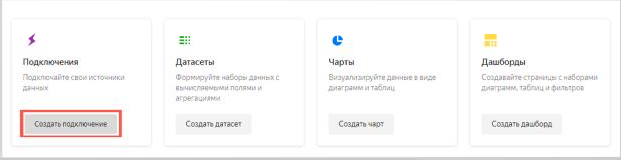
# Новый пользователь Уже использую Yandex Cloud

1. Войдите ( <https://passport.yandex.ru/auth> ) в ваш аккаунт на Яндексе. Если у вас еще нет аккаунта, создайте его (<https://yandex.ru/support/passport/authorization/registration.html> .
2. Откройте главную страницу DataLens (<https://datalens.yandex.ru/> ) .
3. Нажмите **Открыть DataLens**.
4. Нажмите **Войти**.

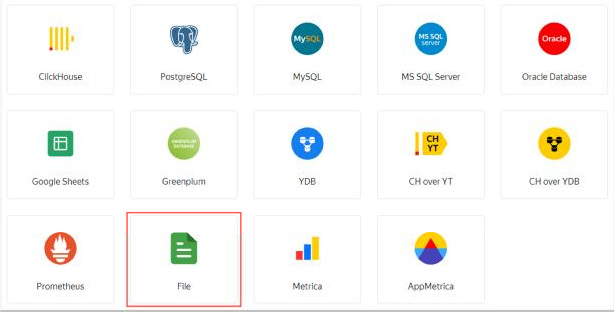
# Шаг 1. Создайте подключение и датасет

Создайте датасет на основе подключения типа **File (**<https://cloud.yandex.ru/docs/datalens/concepts/connection> ) .

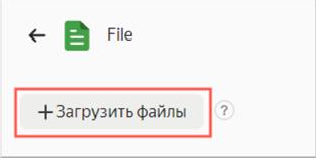
* 1. Перейдите на главную страницу DataLens (<https://datalens.yandex.ru/> ) .
  2. Нажмите кнопку **Создать подключение**.



* 1. Выберите **File**.



* 1. Нажмите кнопку **Загрузить файлы**.

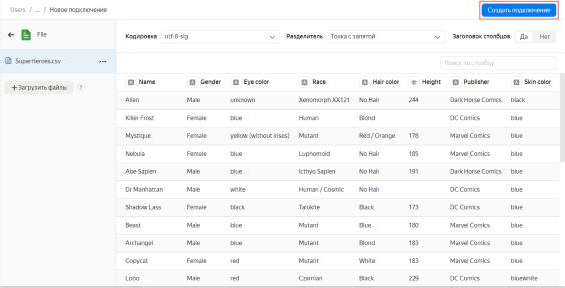


* 1. Выберите необходимый файл. В данном

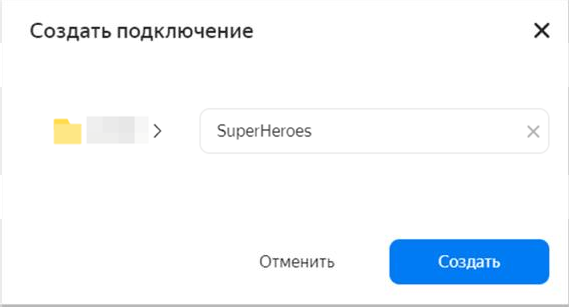
примере **SuperHeroes.csv** (<https://storage.yandexcloud.net/datalens/Super> Heroes.csv ).

Дождитесь, когда на экране появится содержимое таблицы.

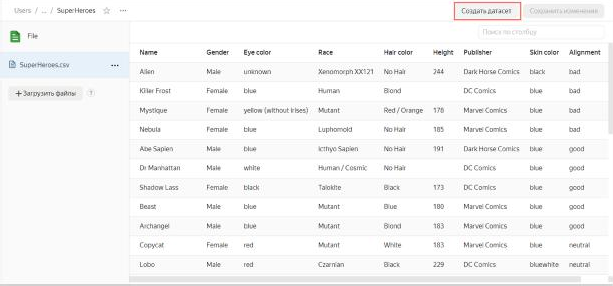
* 1. Нажмите кнопку **Создать подключение**.



* 1. Введите имя подключения SuperHeroes и нажмите **Создать**.

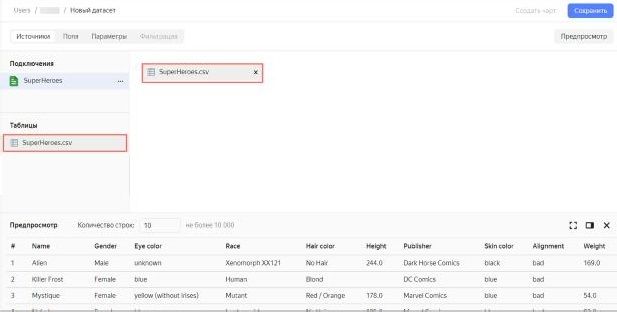


* 1. После сохранения нажмите кнопку **Создать датасет**.



# Шаг 2. Создайте датасет

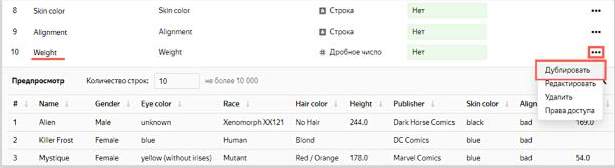
1. Если в рабочей области нет таблицы **SuperHeroes.csv**, то перетащите таблицу из панели выбора в рабочую область.



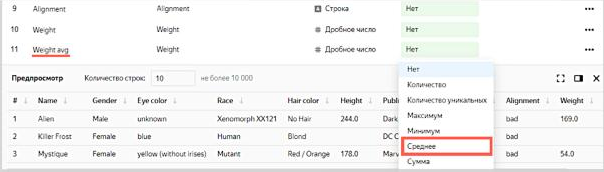
1. Перейдите на вкладку **Поля**.



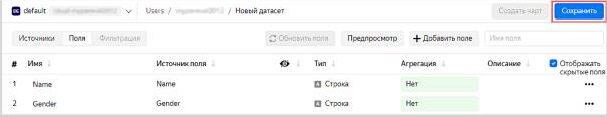
1. Создайте поле с показателем среднего веса супергероя:
   1. Нажмите значок в строке **Weight**.
   2. Выберите **Дублировать**.



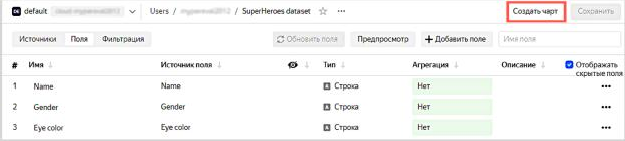
* 1. Переименуйте дубликат поля **Weight (1)** в **Weight avg**: нажмите на имя строки, удалите текущее имя и введите новое.
  2. В столбце **Агрегация** для поля **Weight avg** выберите **Среднее**.



1. В правом верхнем углу нажмите кнопку **Сохранить**.



1. Введите имя датасета **SuperHeroes dataset**, нажмите **Создать**.
2. После сохранения датасета нажмите **Создать чарт**.



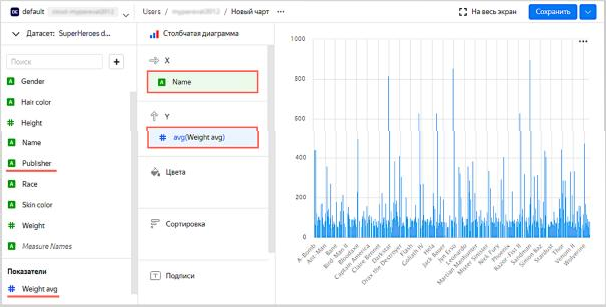
# Шаг 3. Создайте первый чарт

Для визуализации разделения по половому признаку создайте чарт —

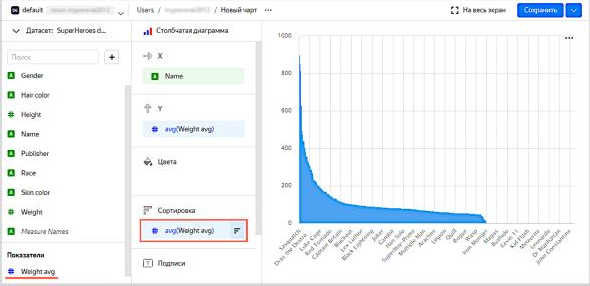
столбчатую диаграмму.

1. Добавьте на график имена супергероев. Для этого из раздела **Измерения** перетащите поле **Name** в секцию **X**.
2. Добавьте на график веса супергероев. Для этого из

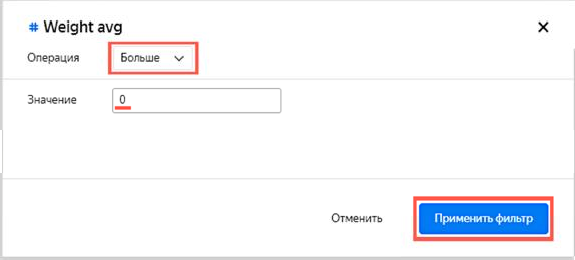
раздела **Показатели** перетащите поле **Weight avg** в секцию **Y**.



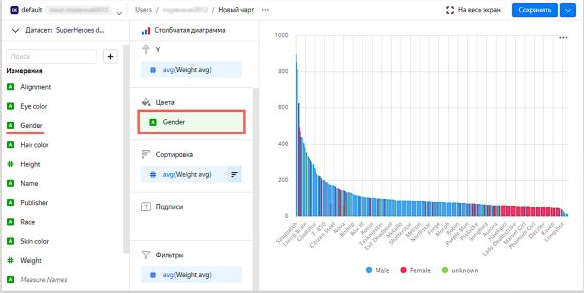
1. Отсортируйте график по весам, например, по убыванию.
   1. Из первой колонки из раздела **Показатели** перетащите поле **Weight avg** в секцию **Сортировка**.



1. Оставьте на графике только тех супергероев, чей вес известен (больше нуля).
   1. Из первой колонки из раздела **Показатели** перетащите поле **Weight avg** в секцию **Фильтры**.
   2. В открывшемся окне укажите операцию **Больше** и значение **0**.
   3. Нажмите кнопку **Применить фильтр**.



1. Добавьте на график цветовое разделение супергероев по половому признаку. Для этого из раздела **Измерения** перетащите поле **Gender** в секцию **Цвета**.

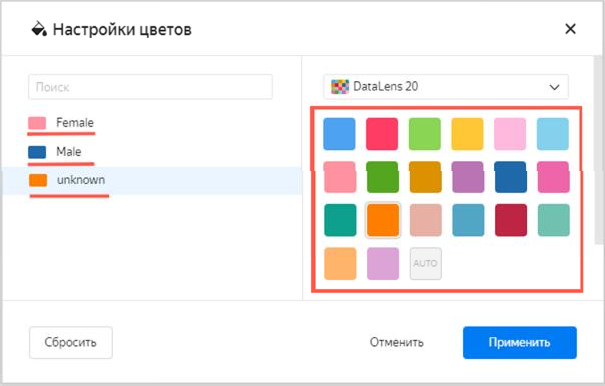


1. Переопределите цвета значений измерения **Gender**.
   1. Нажмите значок настройки в секции **Цвета**.

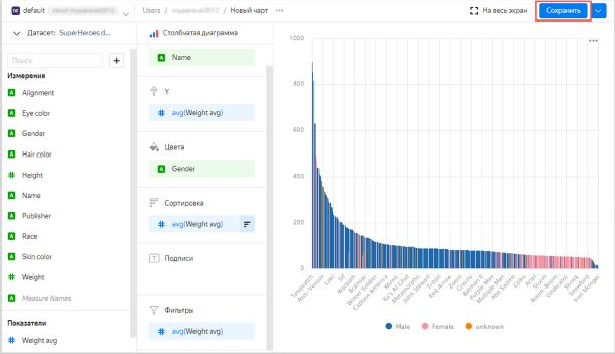


* 1. Выберите цвета: **Female** — розовый, **Male** — синий, **unknown** —

оранжевый.



1. Сохраните чарт.
   1. В правом верхнем углу нажмите кнопку **Сохранить**.

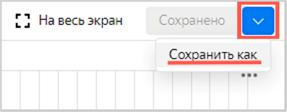


* 1. В открывшемся окне введите название чарта **SuperHeroes — gender** и нажмите кнопку **Сохранить**.

# Шаг 4. Создайте второй чарт

Для визуализации по принадлежности к лагерю хороших или плохих создайте чарт — столбчатую диаграмму.

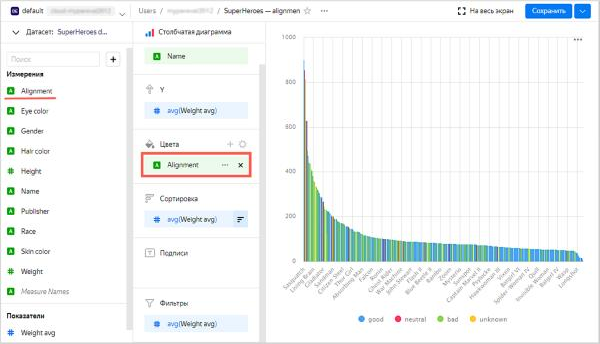
1. Скопируйте чарт, получившийся на предыдущем шаге.
   1. Нажмите значок галочки рядом с кнопкой **Сохранить** в верхнем правом углу.
   2. Нажмите **Сохранить как**.



* 1. В открывшемся окне введите название нового чарта **SuperHeroes — alignment**.
  2. Нажмите кнопку **Готово**.

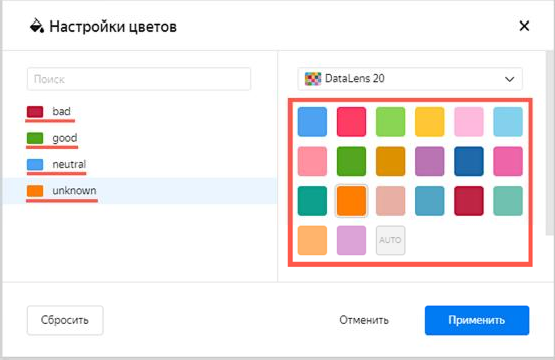
1. Добавьте на график цветовое разделение супергероев по принадлежности к лагерю хороших или плохих. Для этого из раздела **Измерения** перетащите поле **Alignment** в секцию **Цвета**.

Предыдущее значение секции (поле **Gender**) будет заменено на **Alignment**.



1. Переопределите цвета значений измерения **Alignment**.
   1. Нажмите значок настройки в секции **Цвета**.
   2. Выберите цвета: **good** — зеленый, **neutral** — синий, **bad** —

красный, **unknown** — оранжевый.



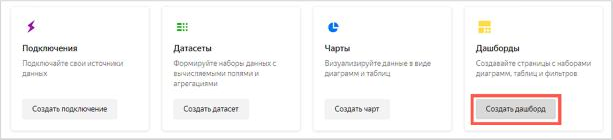
1. В правом верхнем углу нажмите кнопку **Сохранить**.

# Шаг 5. Создайте дашборд

Создайте дашборд, на который будут добавлены чарты

(<https://cloud.yandex.ru/docs/datalens/concepts/dashboard> ) .

1. Перейдите на главную страницу DataLens (<https://datalens.yandex.ru/> ) .
2. Нажмите кнопку **Создать дашборд**.



1. Введите название дашборда **SuperHeroes dashboard** и нажмите кнопку **Создать**.

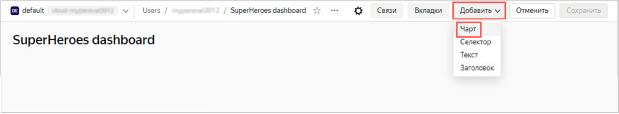
# Шаг 6. Добавьте чарты на дашборд

1. При первом открытии после сохранения дашборд открывается в режиме редактирования. Если вы открыли его повторно, то

нажмите **Редактировать** в правом верхнем углу.



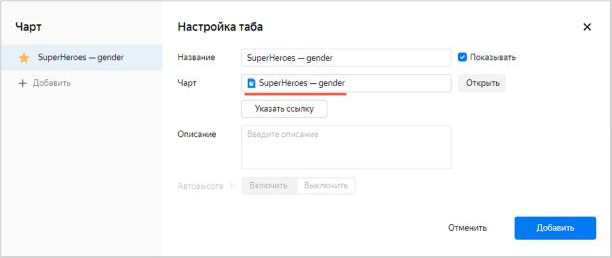
1. Нажмите кнопку **Добавить**.
2. Выберите **Чарт**.



1. В открывшемся окне нажмите кнопку **Выбрать** и выберите чарт **SuperHeroes — gender**.

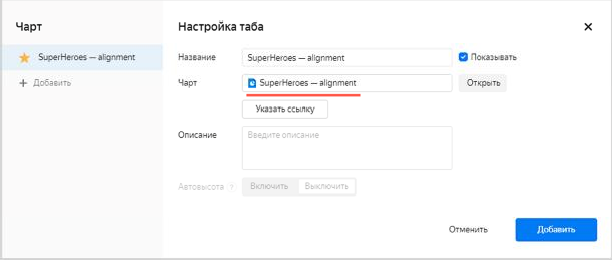
После этого автоматически заполнится поле **Название**.

1. Нажмите кнопку **Добавить**.



1. Добавьте еще один чарт. Пройдите шаг 6 сначала и укажите в пункте 4

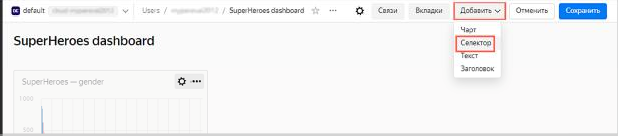
чарт **SuperHeroes — alignment**.



# Шаг 7. Добавьте селекторы на дашборд

Добавьте селекторы (<https://cloud.yandex.ru/docs/datalens/concepts/dashboard#selector> ) , чтобы иметь возможность фильтровать супергероев по расам (поле **Race**) и издательствам (поле **Publisher**).

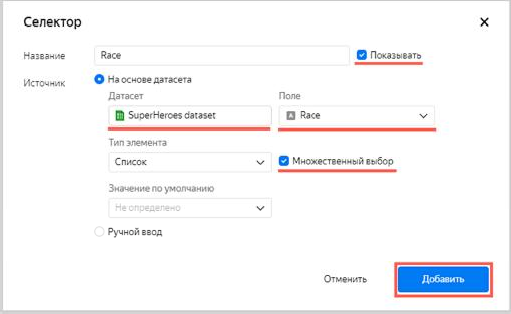
1. Нажмите кнопку **Добавить**.
2. Выберите **Селектор**.



1. Выберите датасет **SuperHeroes dataset**.
2. Выберите поле **Race**.

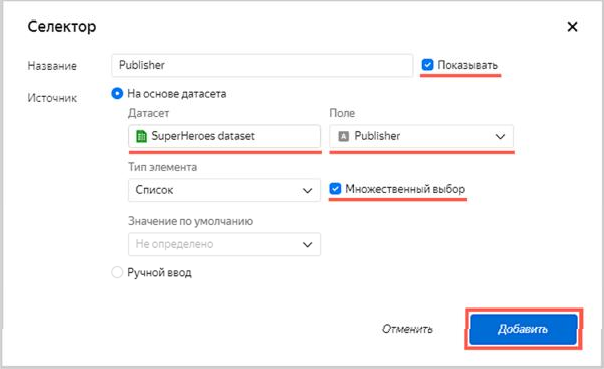
После этого автоматически заполнится поле **Название**.

1. Нажмите галочку **Показывать** напротив заголовка селектора.
2. Включите опцию **Множественный выбор**.
3. Нажмите кнопку **Добавить**.



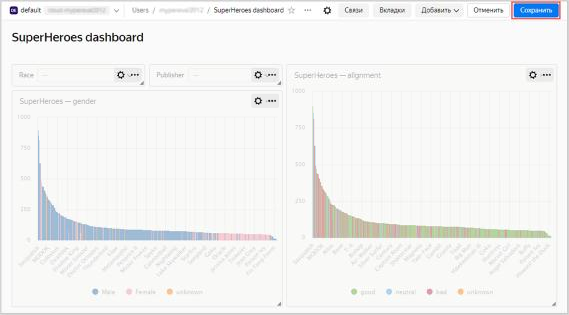
1. Добавьте еще один селектор. Пройдите шаг 7 сначала и укажите в пункте 4

поле **Publisher**.

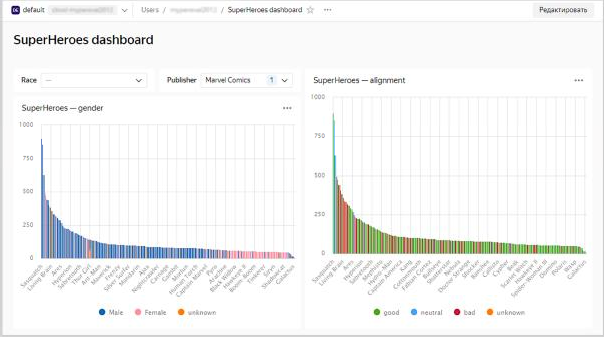


# Шаг 8. Настройте отображение виджетов и приступите к исследованию зависимостей

1. Перетащите селекторы наверх страницы рядом друг с другом.
2. Растяните чарты по ширине и расположите рядом друг с другом.
3. В правом верхнем углу нажмите кнопку **Сохранить**.



1. Примените различные фильтры и исследуйте зависимости веса от пола и лагеря супергероя.



В датасете, который был исследован, прослеживаются следующие зависимости:

* + Мужчины в основном тяжелее женщин.
  + Плохие супергерои в основном тяжелее хороших.

# Шаг 9. Продолжите исследование данных о супергероях

Вы можете создать новые показатели в датасете, например, средний рост (среднее от поля **Height**) и количество супергероев (количество уникальных от поля **Name**) и ответить на вопросы:

* Представителей какой расы больше всего?
* Есть ли зависимость роста от принадлежности к лагерю хороших или плохих?
* Какая студия создала больше всего супергероев?