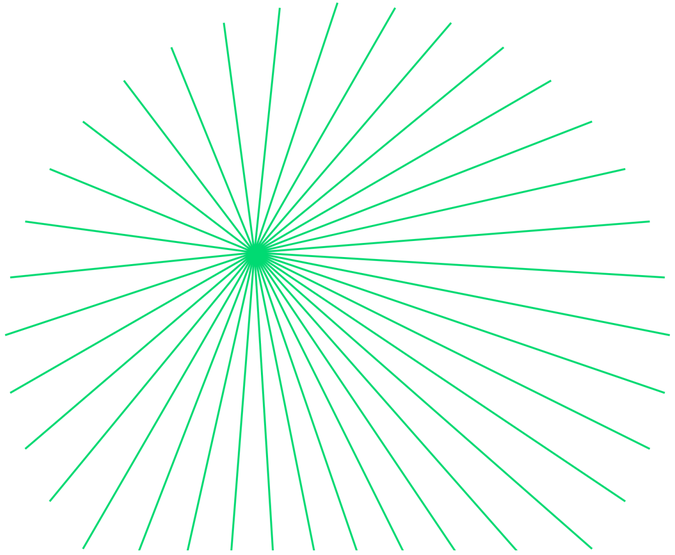
Общая схема проведения анализа данных.

Как измеряется то, что хотим проанализировать.

Вместо предисловия

Сотри случайные черты
И ты увидишь, мир прекрасен
Александр Блок

Аналитика всегда была и всегда использовалась в работе любой компании во все времена. Но как современная профессия стала необычайно популярной именно в последние лет 10-15 лет. Причиной этому то, что бизнес накопил много данных о себе и готов их использовать для улучшения своих процессов и конечных результатов. При этом и компьютерные технологии не стоят на месте, открывая новые возможности в IT отрасли.



Поэтому, и типов аналитиков становится всё больше: продуктовый аналитик, медицинский аналитик, промышленный аналитик, маркетинговый аналитик, бизнес-аналитик, программист-аналитик, аналитик-статистик, вирусный аналитик (занимается анализом данных, связанных с кибербезопасностью, вирусами и способами противодействия им). И со временем сфер, где может понадобиться специалист подобного профиля, будет становиться только больше.

Чем оперирует аналитик

Данные – факты реальной жизни.

Информация – обработанные данные, используемые для принятия решений и решения задач.

Знания – обработанная информация, используемая для принятия решений, решения задач и создания новой информации.

Принятия решений - это процесс, в результате которого ставится проблема (проблемная ситуация), которая снимается за ряд этапов, включая практические действия по устранению проблемной ситуации (реализация найденного решения).

Данные

Без данных вы просто еще один человек с собственным мнением.
Уильям Эдвардс Деминг

Данные это набор значений об изучаемом явлении, ничего не говорящий о причинах и действиях. Например, название города, телефонный номер, название продукта или имя человека - это данные, которые без конкретной цели применения в каком-либо контексте не служат основой для принятия решения. Данные могут быть набором фактов, хранящихся где-то физически, как бумага, как электронное устройство (жесткий диск, флэшка...) или мозг человека.

Предположим, есть компания «Счастливый хомяк», которая специализируется на интернет-продажах корма для грызунов. Отдел маркетинга компании объединил постоянных клиентов в сообщество, организовав Интернет-ресурс. И каждый вход клиента в него фиксируется.

Номер клиента	Данные в базе (вход в ИБ)
111852	1 января 2016
111852	1 января 2016
311853	2 января 2016
211854	3 января 2016
161855	6 января 2016
161855	6 января 2016
161855	6 января 2016
713859	10 января 2016

В табличке выше показано, что каждая строка, это вход клиента в информационную базу. Видно, что некоторые клиента входили несколько раз в день.

Как правило, данные хранятся в базе данных. А когда мы начинаем обрабатывать данные, то они превращаются в информацию. Посмотрим дальше об этом подробнее.

Информация

Появляется в результате обработки данных при решении конкретных задач. И, как писали выше, это результат преобразования данных. Т.е. в базе данных сохраняются именно данные, а не информация. Но когда к базе данных мы отправляем определенный запрос, то получаем по запросу уже требуемую информацию, а не данные.

Продолжим пример с компанией «Счастливый хомяк». Посмотрите на картинку ниже.

Номер клиента	Данные в базе (вход в ИБ)
111852	1 января 2016
111852	1 января 2016
311853	2 января 2016
211854	3 января 2016
161855	6 января 2016
161855	6 января 2016
161855	6 января 2016
713859	10 января 2016

Информация 1

Дата	Вход в ИБ
01.01.2016	2
02.01.2016	1
03.01.2016	1
04.01.2016	0
05.01.2016	
06.01.2016	

Информация 2

Дата	Вход в ИБ
01.01.2016	Да
02.01.2016	Да
03.01.2016	Да
04.01.2016	Нет
05.01.2016	Нет
06.01.2016	Да

Информация 3

Дата	Вход в ИБ (всего)
Январь 2016	25
Февраль 2016	36
Март 2016	33

Информация 4

Дата	Вход в ИБ (среднДн.)
Январь 2016	0,8
Февраль 2016	1,28
Март 2016	1,06

Исходная таблица с данными преобразована в четыре новые таблички.

- Таблица «Информация 1»: Мы сделали запрос в таком виде «Сколько заходов в день?». И был произведён подсчёт количества ежедневного захода клиентов на Интернет-ресурс.

- Таблица «Информация 2»: Мы сделали запрос в таком виде «Каждый ли день был заход на Интернет-ресурс?». И получили табличку, где обозначено был ли хоть один заход на ресурс или не был.

- Таблица «Информация 3»: Тут сделан был такой запрос «Сколько заходов в месяц?». И был произведён суммарный подсчёт количества заходов на ресурс за месяц.

- Таблица «Информация 4»: Произвели расчёт среднего значения захода клиентов в день по месяцам по соответствующему запросу.

Информация

Что объединяет все новые образованные таблицы с информацией? Общее у них одно – над исходными данными был произведён интеллектуальный труд. Можно сказать, что информация – это коммуникация данных и интеллекта.

Номер клиента	Данные в базе (вход в ИБ)
111852	1 января 2016
111852	1 января 2016
311853	2 января 2016
211854	3 января 2016
161855	6 января 2016
161855	6 января 2016
161855	6 января 2016
713859	10 января 2016

Информация 1

Дата	Вход в ИБ
01.01.2016	2
02.01.2016	1
03.01.2016	1
04.01.2016	0
05.01.2016	
06.01.2016	

Информация 2

Дата	Вход в ИБ
01.01.2016	Да
02.01.2016	Да
03.01.2016	Да
04.01.2016	Нет
05.01.2016	Нет
06.01.2016	Да

Информация 3

Дата	Вход в ИБ (всего)
Январь 2016	25
Февраль 2016	36
Март 2016	33

Информация 4

Дата	Вход в ИБ (среднДн.)
Январь 2016	0,8
Февраль 2016	1,28
Март 2016	1,06

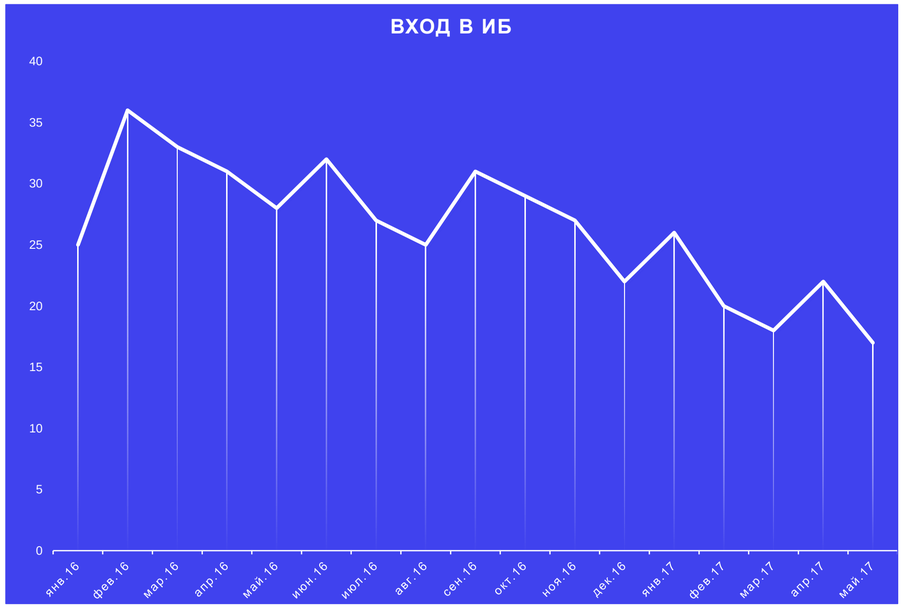
Причём, результат коммуникации способен влиять на суждения о ценности и будущее поведение того, кто воспринимает его. Это происходит, в том числе, и потому, что через такую коммуникацию рождается новое знание.

Знания

Знания связаны с данными, основываются на них, но представляют результат мыслительной деятельности человека, обобщают его опыт, полученный в ходе выполнения какой-либо практической деятельности.

Продолжим рассматривать пример с «Счастливый хомяк». И посмотрим подробно табличку «Информация 3».

Для наглядности она представлена графически (кстати, кто-то из коллег сказал: «Графика – королева аналитики». И в других темах этого курса мы не раз получим доказательства этому).



Что можем заметить? Помесячная динамика заходов на Интернет-ресурс убывающая. Иными словами – «всё нормально, командир, мы падаем». И если менеджмент компании не предпримет чего-нибудь, то активность сообщества приблизится к 0.

Т.е. исходные данные были преобразованы, отображены графически и мы теперь можем делать выводы, т.е. приобретать знания.

Чем оперирует аналитик

Алиса: **Подскажите, пожалуйста, куда мне отсюда идти?**
***Чеширский кот*: Это зависит от того, куда ты хочешь попасть**
Льюис Кэрролл. «Алиса в Стране чудес»

Что конкретно нужно сделать бизнесу, чтобы исправить отрицательную негативную динамику на предыдущем графике?

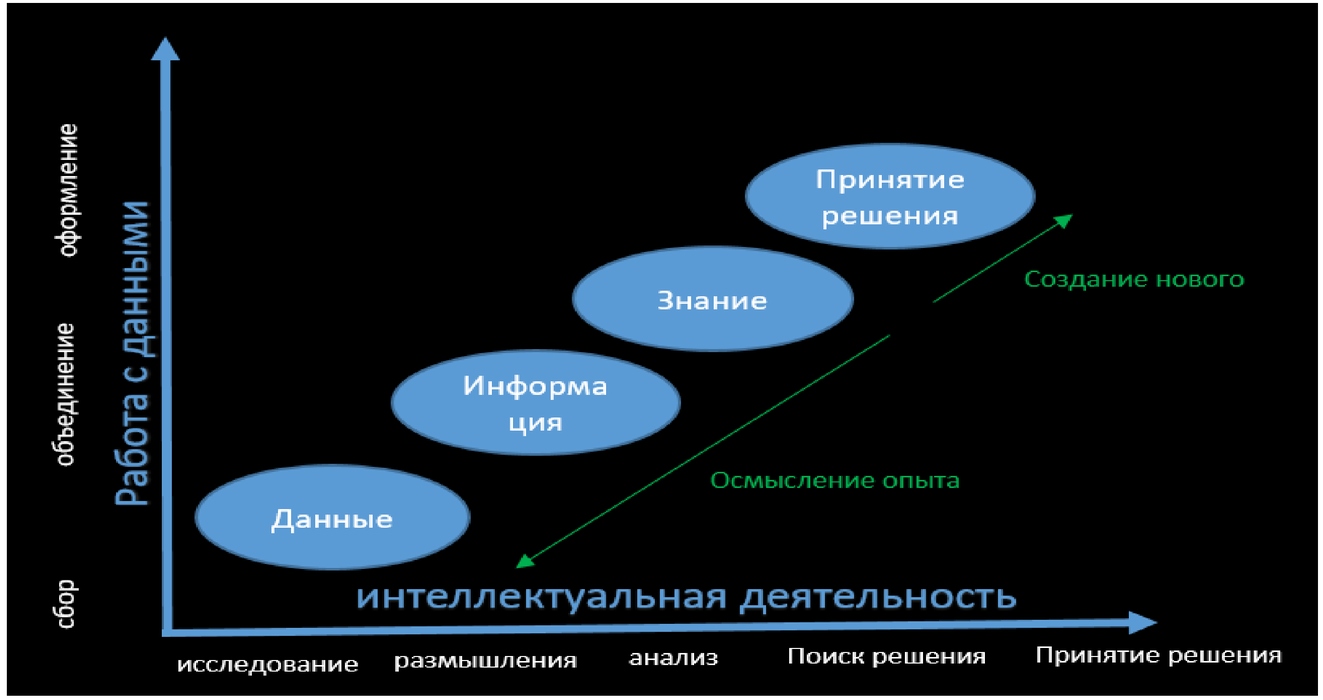
- *Или изменить ценовую политику,*
- *Или изменить сам продукт для клиентов (интернет ресурс из примера выше – это тоже продукт в широком смысле слова),*
- *Или изменить систему продаж,*
- *Или система мотивации сотрудников требует корректировки?*

Ответы на эти вопросы и есть «принятие решения». Иными словами, это процесс, в результате которого ставится проблема (проблемная ситуация), которая решается за ряд этапов, включая практические действия по устранению проблемной ситуации (реализация найденного решения).

В основе каждого успешного решения лежит грамотная аналитическая подготовка.

Отличия и сходства аналитических специальностей

Зафиксируем всё сказанное ранее об аналитической деятельности в общей схеме:



Можно сказать, что эта схема подходит для любой аналитической специальности.

В этом их сходство. Тогда в чём различия специальностей?

Отличия и сходства аналитических специальностей

Метрика — показатель для оценки работы маркетинга, продаж, состояния продукта и т .д. или бизнеса в целом. С помощью метрик можно понять, например, приносит ли прибыль реклама, сколько новых клиентов получила компания за последний месяц, какой процент покупателей совершает повторные покупки.

Специализация	Метрика	Определение
Метрики веб-аналитики сайта (продукта)	CR (Conversion Rate — коэффициент конверсии)	метрика, вычисляемая в процентах, показывает, сколько пользователей совершили конверсионное действие: зарегистрировались, подписались на рассылку, отправили форму, позвонили или заказали обратный звонок, совершили покупку.
	LCR (Lead Close Rate — коэффициент закрытия Лидов в продажу)	метрика, которая показывает долю покупателей от общего числа Лидов. Отражает, насколько хорошо в компании выстроена работа с обращениями и заявками в успешное закрытие сдел
	Sales (продажи)	метрика показывает кол продаж, которые пользо совершили на сайте или последствии после посе сайта.
Финансовые метрики	LTV (Lifetime Value — пожизненная ценность клиента)	етрика показывает всю прибыль от клиента, за весь период его взаимодействия с бизнесом. Есть простые и более сложные формулы расчета пожизненной ценности клиента. Рассмотрим все по порядку
	AC (Customer Acquisition Cost — стоимость привлечения клиентов)	метрика показывает общие затраты на привлечение клиента. Отличается от CPS (Cost per Sale), CPO (Cost per Order), COS (Cost of Sale), тем, что здесь учтены вообще все затраты на привлечение, не только рекламные
Метрики SEO-продвижения	Позиции (ранжирование) ключевых фраз	метрика очень важна, показывает положительную или отрицательную динамику продвижения сайта по целевым запросам, что является основой SEO-продвижения.
	CTR (Click-Through Rate — коэффициент кликабельности)	метрика показывает отношение числа кликов по рекламному объявлению к числу показов. Используется для оценки привлекательности объявления. Оценивать по этому показателю лучше всего не в целом по типу рекламы, а по кампаниям, площадкам показа объявления, типам таргетинга, креативам и т.п.
	CPC (Cost per Click — стоимость клика) или (PPC Pay per Click — оплата за клик)	метрика показывает стоимость за клик по объявлению и является одной из самых распространенных моделей оплаты за рекламу у рекламных площадок. Стоимость клика в первую очередь будет рассчитываться в зависимости от объема конкуренции и только потом от релевантности объявления к посадочной страницы и т.п.
Управление производством	Длительность производственного цикла (ДПЦ)	полное время длительности производственного процесса. Это время, которое необходимо для выполнения всех производственных процессов и циклов производства, до получения готового продукта.
	Коэффициент общей эффективности оборудования (Overall Equipment Effectiveness, OEE)	показатель общей эффективности оборудования и признана в качестве ключевого показателя эффективности оборудования в ряде отраслей
	Время переналадки	метрика является замером времени, которое уходит на переналадку производственного оборудования при переходе с одного производственного процесса на другой

Академия Яндекса

Какие данные бывают

Чтобы анализировать данные, с ними проводят различные математические операции. Но не со всеми можно совершать одинаковые математический действия. В этом разделе рассмотрим, какие типы данных различают, чтобы грамотно с ними потом работать.

Данные делят на два типа – количественные и категориальные:

- Количественные, или числовые. К ним относится всё, что можно измерить числом. Этот тип не делится на другие.

- Качественные, или категориальные. Принимают одно из ограниченного числа фиксированных значений. Они делятся на номинальные, порядковые и бинарные, к каждым их этим типов нужен свой подход.

Номинальные	Порядковые	Количественные
✓ Пол	✓ Оценка в школе	✓ Выручка
✓ Тип товара	✓ Рейтинг товара	✓ Доля рынка
✓ Регион	✓ Уровень образования	✓ Рентабельность
✓ Фамилия	✓ Возрастная группа	✓ Процент брака
✓ Марка машины	✓ Степень ущерба	✓ Количество человек
✓ Удовлетворённость	✓ Оценка рекламы	✓ Возраст

Замечание.
Номинальные двухуровневые данные (пол: м/ж – 2 уровня) часто называют **бинарными**

Какие данные бывают

Давайте поупражняемся и определим тип шкалы () данных в таблицах «Информация...» ниже. Столбец «Дата» – это столбец времени. С ним понятно. Поэтому, рассмотрим только столбец «Вход в ИБ»:

Номер клиента	Данные в базе (вход в ИБ)
111852	1 января 2016
111852	1 января 2016
311853	2 января 2016
211854	3 января 2016
161855	6 января 2016
161855	6 января 2016
161855	6 января 2016
713859	10 января 2016

Информация 1

Дата	Вход в ИБ
01.01.2016	2
02.01.2016	1
03.01.2016	1
04.01.2016	0
05.01.2016	
06.01.2016	

Информация 2

Дата	Вход в ИБ
01.01.2016	Да
02.01.2016	Да
03.01.2016	Да
04.01.2016	Нет
05.01.2016	Нет
06.01.2016	Да

Информация 3

Дата	Вход в ИБ (всего)
Январь 2016	25
Февраль 2016	36
Март 2016	33

Информация 4

Дата	Вход в ИБ (среднДн.)
Январь 2016	0,8
Февраль 2016	1,28
Март 2016	1,06

Какие данные бывают

Номер клиента	Данные в базе (вход в ИБ)
111852	1 января 2016
111852	1 января 2016
311853	2 января 2016
211854	3 января 2016
161855	6 января 2016
161855	6 января 2016
161855	6 января 2016
713859	10 января 2016

Информация 1

Дата	Вход в ИБ
01.01.2016	2
02.01.2016	1
03.01.2016	1
04.01.2016	0
05.01.2016	
06.01.2016	

Информация 2

Дата	Вход в ИБ
01.01.2016	Да
02.01.2016	Да
03.01.2016	Да
04.01.2016	Нет
05.01.2016	Нет
06.01.2016	Да

Информация 3

Дата	Вход в ИБ (всего)
Январь 2016	25
Февраль 2016	36
Март 2016	33

Информация 4

Дата	Вход в ИБ (среднДн.)
Январь 2016	0,8
Февраль 2016	1,28
Март 2016	1,06

«Информация 1»: тип шкалы – количественная. Т.к. посчитали количество заходов на сайт.

«Информация 2»: тип шкалы – бинарная (номинальная 2-х уровневая).

«Информация 3»: тип шкалы – количественная. Т.к. посчитали количество заходов на сайт.

«Информация 4»: тип шкалы – количественная. Т.к. посчитали среднее количество заходов на сайт.

Начало аналитической работы в питоне

**Считайте, что поддается подсчету, измеряйте,
что поддается измерениям,
а не измеряемое делайте измеряемым.**

Галилео Галилей

Обычный план аналитического проекта

1. Проведение обзора данных (EDA)

Первичное исследование данных:
- Импорт необходимых библиотек;
- Чтение файлов и сохранение полученных данных в переменные;
- Получение общей информации о таблицах (head, info, describe);
- Графическое представление данных из таблиц
- Выводы

Предобработка данных:
.....

Какие методы могут понадобиться в работе

**Считайте, что поддается подсчету, измеряйте,
что поддается измерениям,
а не измеряемое делайте измеряемым.**
Галилео Галилей

0. Библиотека **Pandas**
1. Загрузка данных: **read_csv()**
2. Просмотр загруженных данных: вывести первые строки **head(n),** последние строки **tail(n)**
3. Работа с таблицами: **qroupby()**

Академия Яндекса

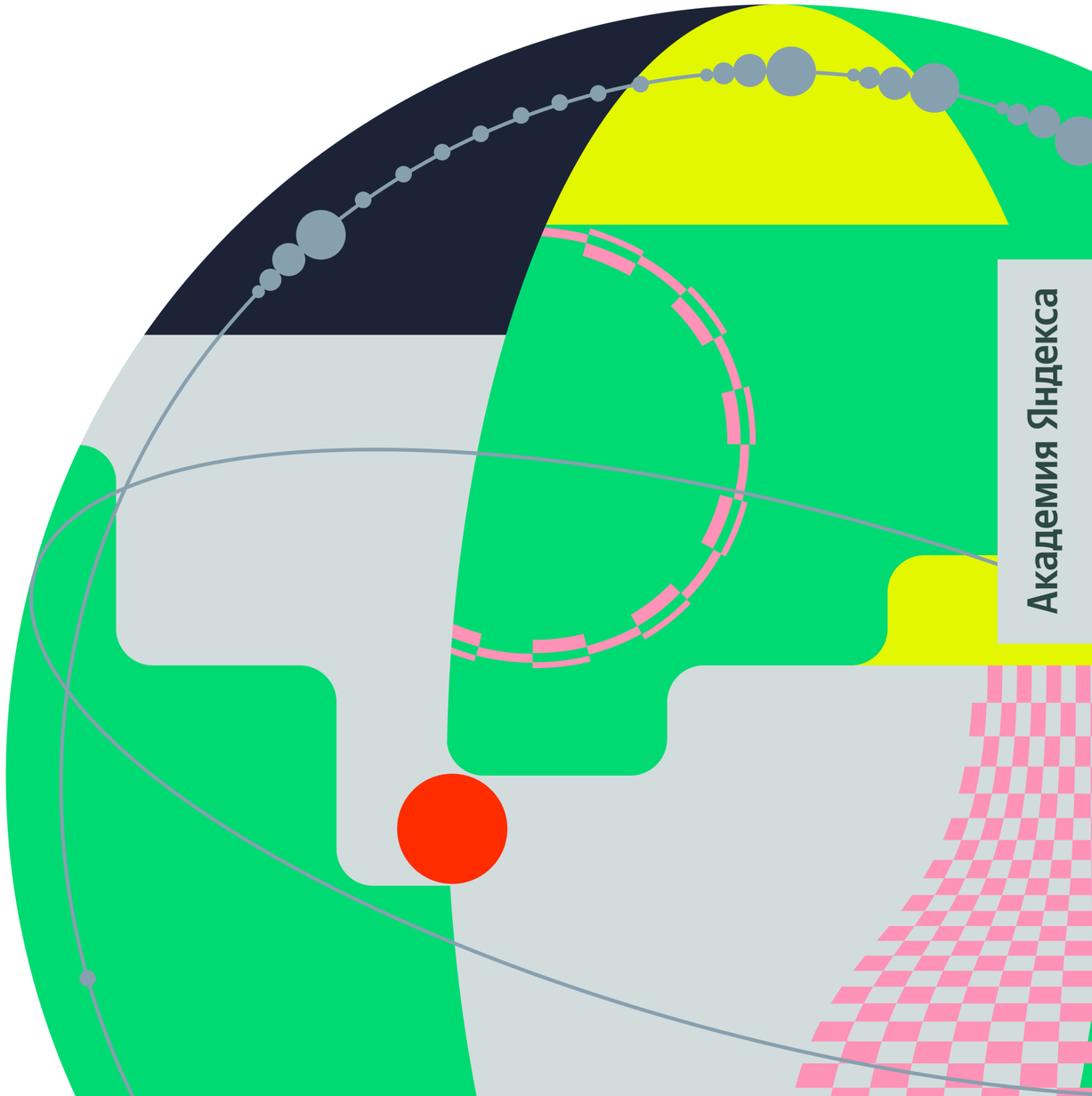
Дополнительные материалы

Разница groupby и pivot_table: https://stackoverflow.com/questions/34702815/difference-between-groupby-and-pivot-table-for-pandas-dataframes

Использование фильтра к данным: https://fullstacker.ru/filtraciya-dannyh-v-pandas-uslovnye-operatory-i-metod-query#использование-eval()-для- фильтраци

Посмотреть каггл: https://www.kaggle.com/code/emstrakhov/eda-with-pandas/notebook

Спасибо за внимание



Академия Яндекса позволяет школьникам и студентам освоить востребованные IT-профессии по программам, разработанным экспертами компании