

Информационно-справочные материалы к модулю 1.

(материалы представлены с сайта <https://wiki.loginom.ru>)

Модуль 1. Технология OLAP

Содержание

1. Существующие программные решения для OLAP – моделирования

- 1.1. Оперативный анализ данных (OnLine Analytical Processing)
- 1.2. Куб (Cube)
- 1.3. Хранилище данных (Data Warehouse)

2. Построение и анализ OLAP-кубов

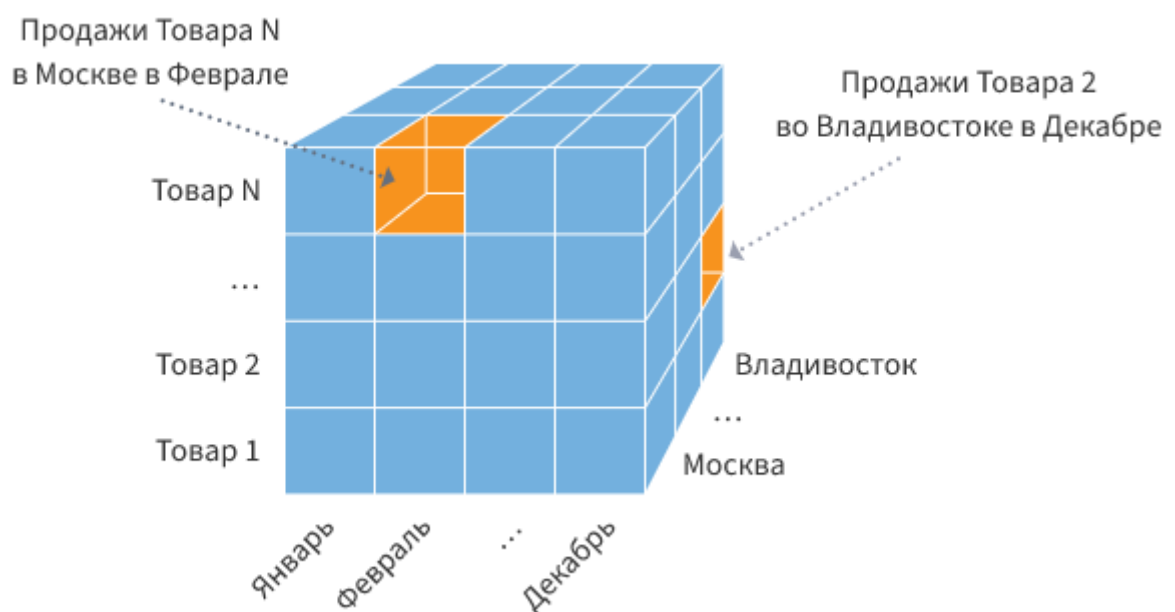
- 2.1. Из бизнес-пользователя в аналитики. Какую систему для анализа данных выбрать?
- 2.2. Анализ ABC (ABC-analysis)
- 2.3. Анализ XYZ (XYZ analysis)
- 2.4. ABC XYZ — анализ продаж для умного управления бизнесом
- 2.5. Анализ RFM (RFM-analysis)

1.Существующие программные решения для OLAP – моделирования

1.1. Оперативный анализ данных (OnLine Analytical Processing)

Оперативный анализ данных - технология хранения и обработки **многомерных данных**, позволяющая получать сложные аналитические отчёты в реальном времени.

В основе технологии лежит представление данных в виде многомерных кубов, где **измерениями** являются категории, а в ячейках внутри куба содержатся **факты** и **агрегаты**.



Автором идеи OLAP является Эдгар Кодд, который сформулировал 12 правил, определивших эту технологию:

1. Многомерный концептуальный взгляд на данные (Multidimensional conceptual view).

2. Прозрачность для пользователя (Transparency).
3. Доступность разнородных источников данных (Accessibility).
4. Постоянство характеристик производительности при увеличении числа измерений (Consistent reporting performance).
5. Клиент-серверная архитектура (Client server architecture).
6. Общность измерений по структуре и возможностям обработки (Generic Dimensionality).
7. Обработка разреженных матриц (Dynamic sparse matrix handling).
8. Наличие многопользовательской среды (Multi-user support).
9. Операции с любым числом измерениями (Unrestricted cross-dimensional operations).
10. Интуитивное манипулирование данными (Intuitive data manipulation).
11. Гибкое формирование отчётности (Flexible reporting).
12. Неограниченное число измерений и уровней агрегирования данных (Unlimited Dimensions and aggregation levels).

В настоящее время список из этих 12 правил расширили до 18 главных правил, а всего их около 300.

Альтернативой приведенным выше правилам для определения OLAP является так называемый **тест FASMI** (Fast Analysis of Shared Multidimensional Information – быстрый анализ разделяемой многомерной информации). Он включает **пять критериев**, которым должно удовлетворять приложение, чтобы относиться к категории OLAP:

- высокая скорость выполнения аналитических запросов,
- мощная подсистемы анализа,

- организация разделенного доступа к данным,
- многомерное представление данных,
- доступность информации.

Есть несколько **разновидностей архитектур OLAP**:

- DOLAP (Desktop OLAP) – настольный OLAP. Продукты для локального многомерного анализа, не поддерживающие многопользовательский режим.
- ROLAP (Relational OLAP) – реляционный OLAP. Системы, в которых многомерность эмулируется с помощью **реляционной СУБД**.
- MOLAP (Multidimensional OLAP) – многомерный OLAP. Обеспечивает максимальную производительность, так как его структура и интерфейс наилучшим образом соответствуют структуре аналитических запросов.
- HOLAP (Hybrid OLAP) – гибридный OLAP. Определяет многомерные инструменты анализа, которые прозрачным для пользователя способом сохраняют данные или в реляционной, или в **многомерной базе данных**.

1.2. Куб (Cube)

Куб представляет собой многомерный массив данных, используемый в системах **оперативной аналитической обработки** (OLAP). Кроме этого, куб можно рассматривать как модель данных, обеспечивающую функциональность OLAP.

В основе идеи построения куба лежит многомерная модель данных, предполагающая их разделение на измерения и факты. Куб, который содержит более чем три измерения, называется многомерным.

Куб можно рассматривать как многомерное обобщение двумерной электронной таблицы. При этом измерения образуют оси многомерной модели данных (ребра куба), а факты — ячейки внутри куба, расположенные на пересечении соответствующих значений измерений.

На рисунке представлен пример 3-мерного куба, содержащего измерения «Товар», «Месяц» и «Город».

OLAP-куб

Представление данных в виде куба обеспечивает возможность реализации концептуально простых операций для поддержки процесса анализа — срез и фрагментацию, детализацию, свертывание и вращение:

Срез (slice) — извлечение из куба подмножества ячеек, связанных с каким-либо значением одного из его измерений. Например, срез по значению «Февраль» измерения «Месяц» (см. рис.) позволяет получить данные по продажам всех товаров во всех городах. Фактически, срез можно рассматривать как одномерный куб, который

можно представить в виде обычной плоской таблицы. Использование срезов позволяет выполнить декомпозицию задачи анализа сложных многомерных структур на несколько более простых одномерных задач.

Фрагментация (dice) — извлечение из куба некоторого подкуба, содержащего только те значения измерений, которые нужны для анализа. Например, фрагмент куба может содержать данные только за определенный интервал дат или о продажах по заданным городам.

Детализация (Drill Down/Up) — позволяет аналитику изменять уровень представления данных в кубе от более общего к более детальному (down) или наоборот (up).

Свертывание (RollUp) — агрегирование данных по одному или нескольким измерениям. Производится в том случае, если нужны сводные, а не полные данные.

Вращение (pivot) — позволяет менять пространственную ориентацию осей измерений куба, выбирая наиболее удобное для аналитика представление.

В OLAP-технологиях куб — это прежде всего средство визуализации многомерных данных. Поэтому при его использовании приходится решать задачу отображения информации в удобном и интерпретируемом для человека виде.

Для представления данных в кубе разработан специальный визуализатор, называемый **кросс-таблицей**. Она представляет данные в виде таблицы, но снабжена специальным интерфейсом, который позволяет оперативно группировать измерения, управлять срезами куба и отображать их на плоскости.

		Средства		Средства		Средства		Средства
		Средства	Средства	Средства	Средства	Средства	Средства	
Средства	Средства							Средства
	Средства							Средства
	Средства							Средства
	Средства							Средства
Средства		Средства	Средства	Средства	Средства	Средства	Средства	Средства
Средства	Средства							Средства
	Средства							Средства
	Средства							Средства
Средства		Средства	Средства	Средства	Средства	Средства	Средства	Средства

Таким образом, кросс-таблица позволяет заменить сложное, трудно интерпретируемое многомерное представление множеством «плоских», но более простых и понятных.

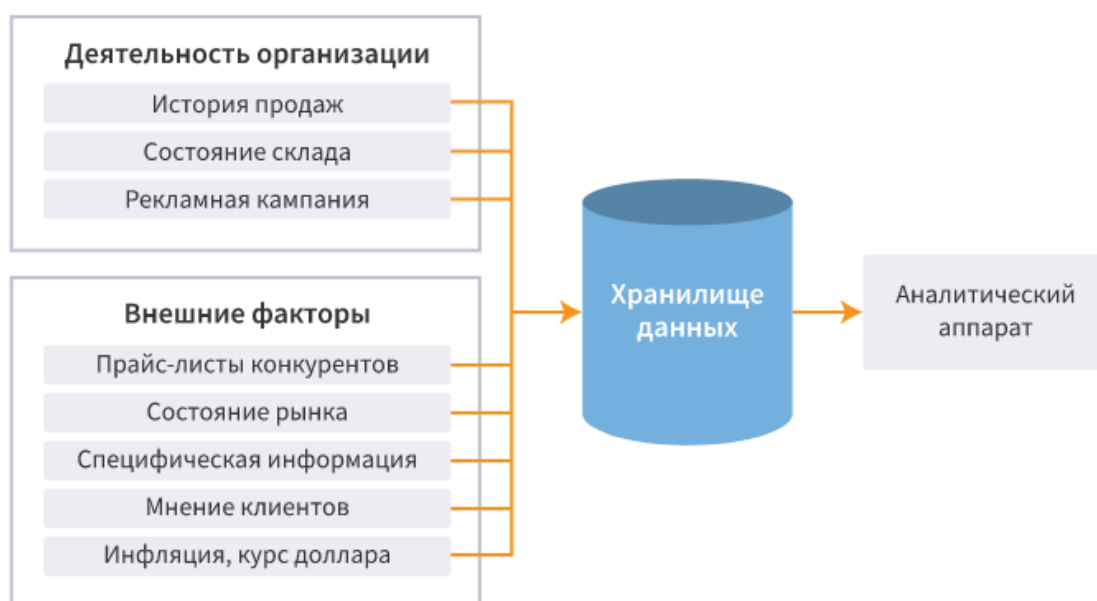
Впервые концепция OLAP-кубов была описана **Эдгаром Коддом**.

1.3. Хранилище данных (Data Warehouse)

Хранилище данных представляет собой предметно-ориентированный, интегрированный, неизменяемый и поддерживающий хронологию набор данных, организованный для целей поддержки принятия решений. Хранилища данных позволяют эффективнее, быстрее и качественнее предоставлять данные для систем их аналитической обработки, чем обычные СУБД.

- **Предметная ориентированность** означает, что данные в хранилище объединены в соответствии с областями, которые они описывают, а не с приложениями, которые их используют.
- **Интегрированность** означает, что хранилище должно поддерживать совместное хранение данных различной природы, форматов и типов, отражающих различные аспекты предметной области, а не отдельные бизнес-функции. Данные содержатся внутри хранилища в его едином внутреннем формате.
- **Неизменчивость** подразумевает, что для данных в хранилище предусмотрена только операция добавления, а удалять или изменять их нельзя. Если какие-либо изменения все же необходимы, «перегружается» все хранилище целиком. Необходимость такого подхода объясняется тем, что при промышленной эксплуатации хранилища совместно с аналитическими платформами один и тот же запрос к нему, выполняемый в любое время, должен обеспечить предоставление одних и тех же данных. Очевидно, что если бы в хранилище были разрешены изменения, то два одинаковых запроса, выполняемые с некоторым интервалом, в течение которого данные могли измениться, сформируют два различных набора данных, анализ которых может привести к некорректным заключениям и выводам, что недопустимо.

Поддержка хронологии указывает на то, что хранение данных организовано с учетом даты и времени их появления, для чего каждой записи присваивается специальная метка времени (time stamp), что позволяет извлекать данные в хронологическом порядке и анализировать временные последовательности.



Хранилища могут использовать реляционную модель, когда данные в них нормализованы, или многомерную, использующую так называемые измерения. В нормализованных хранилищах данные содержатся в таблицах третьей нормальной формы. Преимущество нормализованных ХД заключается в простоте разработки и управления. Недостатком является необходимость денормализации данных «на лету» при их извлечении из множества таблиц при выполнении сложных аналитических запросов.

При формировании больших выборок это приводит к значительным задержкам в получении данных, а если хранилище и аналитическая платформа интегрированы в информационную систему предприятия, то возрастает нагрузка на всю систему, что может осложнить работу многих пользователей. Данную проблему

частично удастся решить, используя в хранилище модель данных, основанную на измерениях. Применяются две разновидности многомерных моделей данных — «звезда» и «снежинка». Все загружаемые в хранилище данные обязательно должны быть определены как измерение, атрибут либо факт.

Кроме собственно данных, описывающих бизнес-процессы компании, в хранилище содержатся метаданные — служебные данные, описывающие структуру хранилища, содержащие информацию о принадлежности данных к тому или иному типу или виду (измерение, атрибут или факт). С помощью метаданных формируется семантический слой, который обеспечивает визуальные средства управления данными и метаданными. Метаданные в хранилищах разделяют на технические (обеспечивают работу самого хранилища), и бизнес-метаданные (описывают структуру данных в рамках заданной бизнес-модели).

В промышленной эксплуатации основными источниками данных для хранилищ являются OLTP-системы. Кроме этого, источниками быть любые файлы в информационной системе предприятия, где содержится структурированная информация, анализ которой, как ожидается, может дать полезные знания. Такие файлы могут иметь различные типы и форматы — электронные таблицы (Excel), настольные СУБД (Access), текст с разделителями (TXT, CSV-файлы), файлы учетных систем (1С:Предприятие, Парус) и т.д. Поэтому для хранилищ данных очень важно иметь развитые средства для загрузки и интегрирования данных из различных типов и форматов.

Автором концепции хранилищ данных в том виде, в каком она существует в настоящее время, считается Билл Инмон, который ввел

данный термин в 1970-х. Большой вклад в развитие теории хранилищ данных и практики их использования в области бизнес-анализа и поддержки принятия решений внес Ральф Кимбалл, который также является автором многомерной модели данных.

2. Построение и анализ OLAP-кубов

2.1. Из бизнес-пользователя в аналитики. Какую систему для анализа данных выбрать?

Хотите начать самостоятельно анализировать данные и не знаете, какой аналитический инструмент подойдёт? Где приручать питонов, а где ограничиться Excel? Изучите обзор различных подходов к анализу данных, их преимуществ и недостатков.

Почему каждый специалист хочет стать аналитиком и кто такой Citizen data scientists?

Под начинающим аналитиком мы в большей степени подразумеваем специалиста с техническим или экономическим образованием, но не исключаем и другую непрофильную специализацию.

В современном мире тяжело представить компанию, которая не собирает данные и не ориентируется при этом на «data-driven»-подход. На фоне информатизации более востребованными становятся специалисты, которые помогают бизнесу принимать решения на основе данных: Data Scientists и аналитики данных. Эти профессии являются самыми высокооплачиваемыми и перспективными специальностями в IT-сфере, а спрос на них продолжает расти.

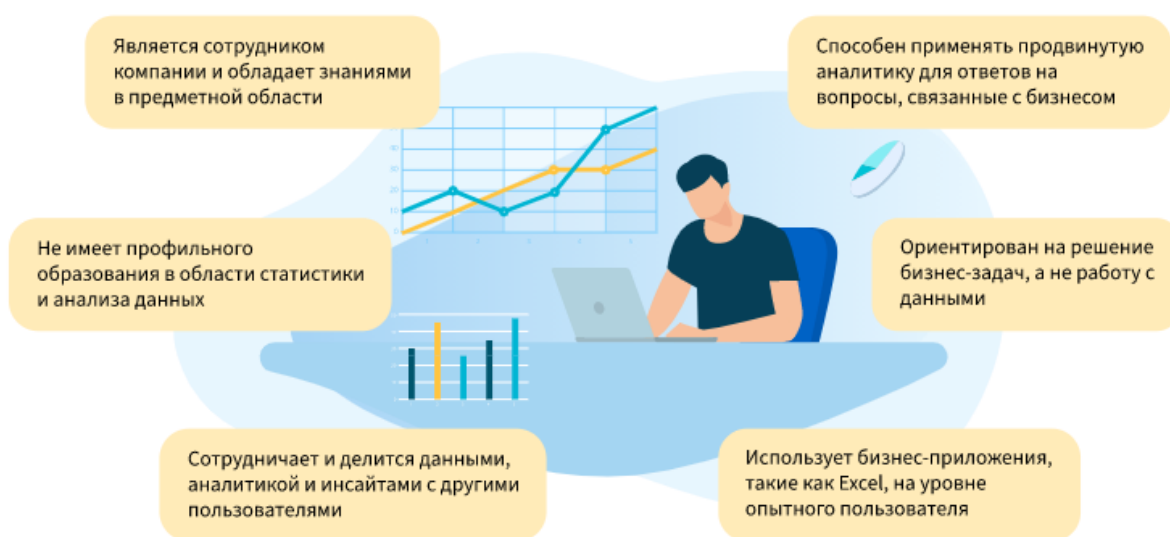
Специалисты в предметных областях постепенно переквалифицируются в аналитиков данных. Связано это с тем, что бизнес-эксперты являются основными носителями ключевых знаний о компании и хотят использовать эти данные. У них есть информация и сотни идей, как улучшить бизнес. Например, маркетологам важно проводить маркетинговые исследования, финансистам — искать зависимость между финансовыми показателями, а специалистам по запасам — прогнозировать спрос на продукцию.

Однако в большинстве компаний весь анализ данных завязан на IT-отделе. Из-за этого возникает ряд проблем:

«Хотелок» у бизнеса много, IT-ресурсов мало, рынок меняется быстро. Возникает боль всех заказчиков — очередь к IT-специалистам. Из-за этого ожидание реализации задачи может длиться месяцами.

Эксперты понимают бизнес, но не понимают язык программистов. Программисты, наоборот, не знают тонкостей бизнеса. Из-за разницы в толковании и терминологии, объяснении постановки задач и обсуждении технического задания срок реализации растягивается.

У бизнес-пользователей всё чаще появляется необходимость анализировать данные собственными силами, проверять гипотезы на практике и получать работающие прототипы систем, быстро решать свои задачи, не дожидаясь разработчиков. Это стремление привело к появлению новой роли в аналитике — гражданский специалист по работе с данными (Citizen data scientist).



Портрет Citizen data scientist

Этот специалист умеет создавать и генерировать модели продвинутой аналитики и прогнозирования. При этом основная его роль

выходит за рамки статистики и аналитики — прежде всего он остается бизнес-экспертом внутри своего подразделения. Citizen data scientists не является профессионалом в области интеллектуального анализа данных и Big Data, у него нет специального образования и глубоких навыков в этой сфере. Зато он привносит в этот процесс собственный опыт и уникальные предметные знания.

Для воплощения своих идей в жизнь гражданскому специалисту по работе с данными требуется подходящее программное обеспечение. Именно развитие технологий послужило ключевым фактором роста числа Citizen data scientists. Аналитические инструменты для неспециалистов стали доступнее в использовании, обеспечивают упрощённую подготовку, обработку данных и расширенную аналитику, включающую в себя Machine Learning и другие инструменты Data Science.

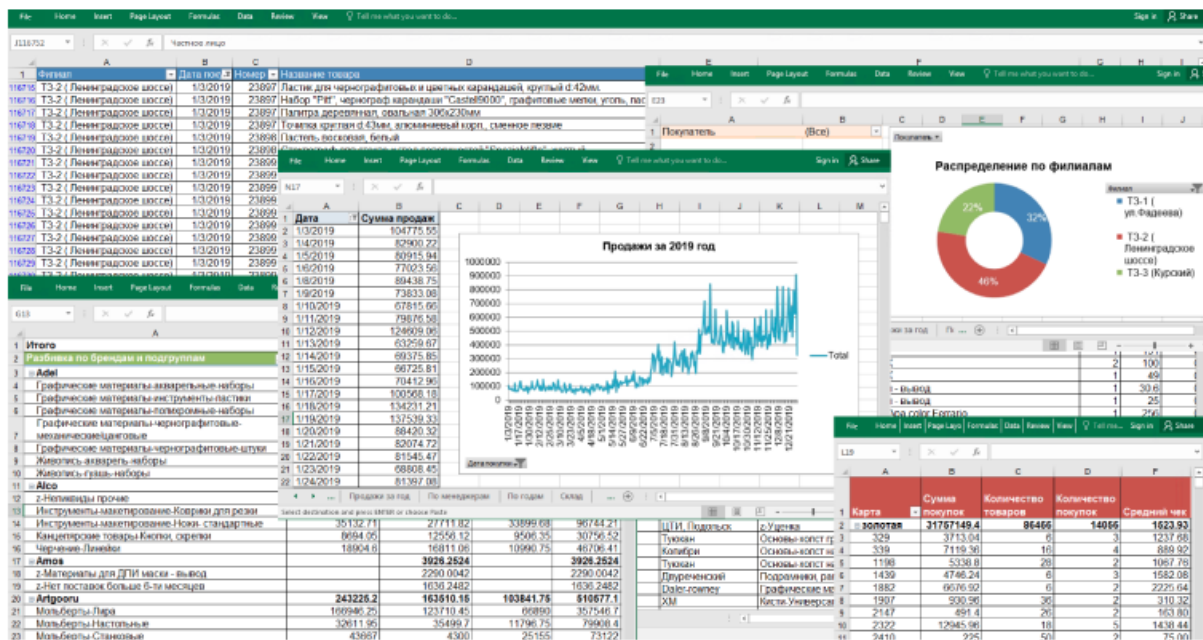
Какой инструмент для анализа данных выбрать начинающему аналитику? Давайте рассмотрим популярные классы систем для анализа данных и их особенности.

Excel

Бытует теорема о полноте Excel: любой бизнес-процесс можно описать достаточно «жирным» excel-файлом. Действительно, этот табличный редактор — настольный и универсальный инструмент любого специалиста по работе с данными. До сих пор ни один инструмент аналитика не может превзойти Excel по популярности.

Быстро произвести разнообразные расчёты, построить сводные таблицы, рассчитать прогноз и показать графики руководству — для этого вполне подходит табличный редактор. Анализ данных в Excel можно выполнить с помощью статистических процедур и функций (корреляция, регрессия, скользящее среднее и т.д.). Есть надстройки и

приложения Microsoft, которые расширяют возможности Excel для очистки данных, создания моделей и отчётов сложной структуры, инструменты визуализации и другие.



Пример работы в Excel

Производительность Excel — недостаток программы, который особенно ощущается при росте объёма данных до одного миллиона строк: система начинает медленно производить вычисления. Иногда из-за этих трудностей с Excel-таблицами становится невозможно работать.

Проблемы с Excel появляются, когда компания растёт, в подготовке одного отчёта в Excel участвуют несколько сотрудников, которые постоянно обмениваются файлами, требуется автоматизация или сложная многоэтапная обработка. Например, каждую неделю разные подразделения готовят отчёты для коммерческого директора, склеивая данные из нескольких Excel-таблиц и выгрузок из 1С, с десятками вкладок и ссылок в нескольких версиях, да ещё постоянно изменяют и «улучшают» эти отчёты. Написание макросов на встроенном в MS Office языке помогает решить проблему, но

ненадолго. В конце концов, компания сталкивается с состоянием, для которого даже есть собственное определение — Excel Hell.

Резюме:

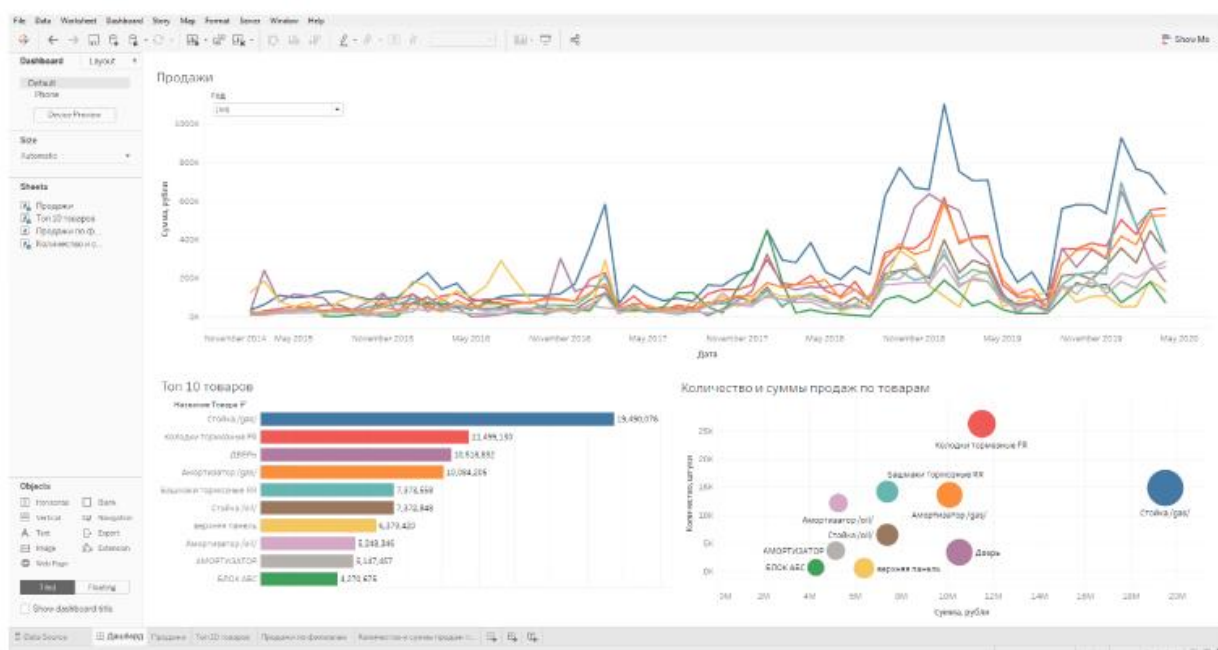
+ : Excel знаком каждому, поэтому подойдёт всем начинающим аналитикам.
Сфера применения: для быстрого индивидуального исследования гипотез на небольшом объёме структурированных данных.

— : Когда танцы с бубном над подготовкой данных и сводными отчётами начинают занимать до нескольких часов в сутки, данных становится много, информационная модель усложняется, с отчётами работает несколько человек, появляется необходимость в изменении бизнес-процессов и переходе на другой инструмент.

BI-системы

Традиционные системы Business Intelligence — удобные инструменты представления и визуализации информации. К ним относятся Power BI, Tableau и другие.

BI-платформы позволяют собирать данные из различных источников, строить регулярные красивые отчёты и интерактивные дашборды для руководителей с любой степенью детализации. Они используются для создания систем аналитической отчётности, мониторинга, KPI и отвечают на вопросы: что случилось ранее или происходит в текущий момент. Эти продукты способны обработать во много раз больше данных, чем Excel.



Пример работы в Tableau

К недостаткам BI-систем относится отсутствие инструментов для продвинутой аналитики (кроме встроенных сторонних языков программирования). Без погружения в кодирование пользователь не сможет заниматься именно анализом и предсказанием развития ситуации в будущем: почему это случилось, что может случиться и что делать. Например, кто из клиентов склонен к оттоку или какие факторы влияют на продажу товаров.

Компаниям преподносят BI-системы как решение проблем получения отчётов. На самом деле BI — это только вершина айсберга, а под водой скрывается множество сложностей получения данных — ETL (Extract, Transform, Load). Загрузка, предобработка, очистка и стандартизация данных — это самая большая проблема аналитиков, которая занимает до 80% всего процесса анализа данных. Для подготовки только одного отчёта на ETL-процесс может уходить до нескольких недель. Например, когда необходимо совместить данные о производстве и поставках, которые вносились разными отделами в разных местах и системах. BI-платформы предлагают инструменты или

дополнительные компоненты для ETL-процесса, но их функционал либо ограничен и недостаточен, либо необходимо писать код.

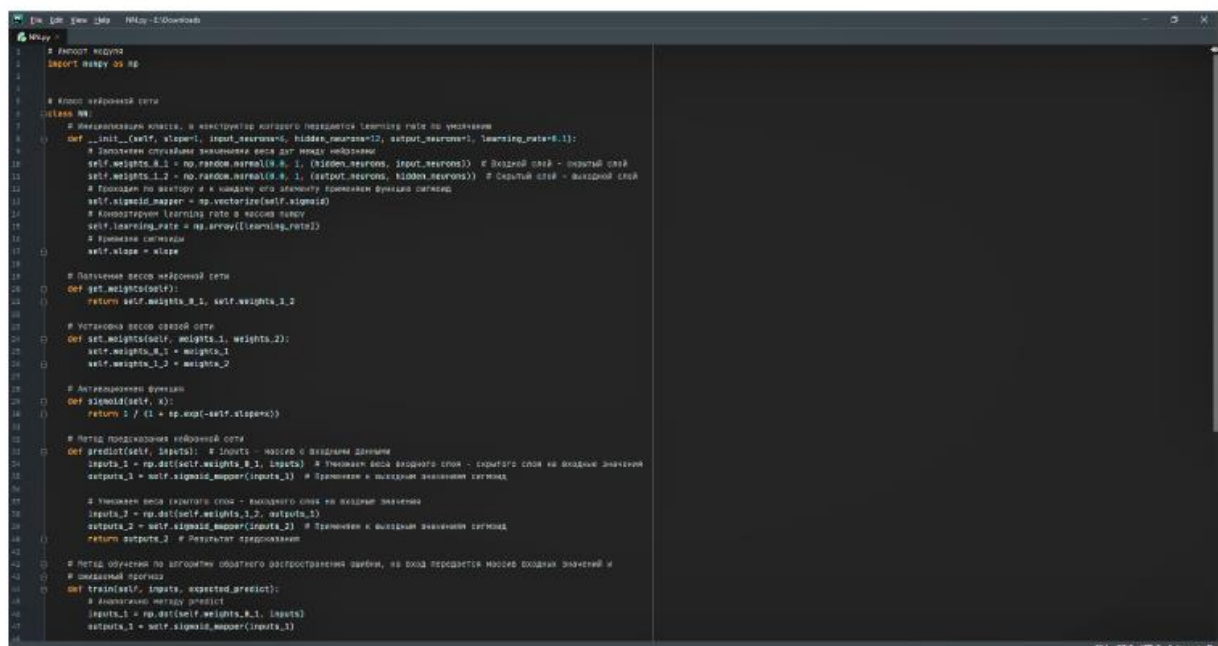
Резюме:

+ : Для бизнес-пользователей погружение в BI-приложения осуществляется легко и просто, оно не требует специальных знаний. С помощью BI-систем можно строить красивые отчёты для руководителей и проводить визуальную оценку для поиска инсайтов.

— : Если в компании не существует единого хранилища данных и не налажены процессы управления качеством информации, то придётся порядком попотеть совместно с IT-шниками над получением достоверных результатов. Для возможностей, связанных с углублённой аналитикой данных, надо использовать другие инструменты.

Языки программирования для анализа данных

Среди аналитиков популярны языки программирования Python и R. Они мощные и гибкие, что позволяет написать на них практически всё, что необходимо, работают с большими объёмами данных. В помощь Data Scientist'ам существует огромное количество готовых библиотек для визуализации, ETL, машинного обучения и интеллектуального анализа данных. Часто появляются новые библиотеки, которые размещаются в открытом доступе.



```
1 # Импорт модуля
2 import numpy as np
3
4 # Класс нейронной сети
5 class NN:
6     # Инициализация класса, в конструкторе задается количество нейронов в каждом слое
7     def __init__(self, input_neurons, hidden_neurons, output_neurons, learning_rate=0.1):
8         # Инициализация случайными значениями веса для каждой нейронной
9         self.weights_1,2 = np.random.randn(4, 1, (hidden_neurons, input_neurons)) # Входной слой - скрытый слой
10        self.weights_2,3 = np.random.randn(4, 1, (output_neurons, hidden_neurons)) # Скрытый слой - выходной слой
11        # Функция по вектору и к каждому его элементу применяется функция сигмоид
12        self.sigmoid_function = np.vectorize(self.sigmoid)
13        # Константа learning_rate в массив numpy
14        self.learning_rate = np.array([learning_rate])
15        # Константа скорости
16        self.clamp = clamp
17
18    # Получение весов нейронной сети
19    def get_weights(self):
20        return self.weights_1,2, self.weights_2,3
21
22    # Установка весов связей сети
23    def set_weights(self, weights_1, weights_2):
24        self.weights_1,2 = weights_1
25        self.weights_2,3 = weights_2
26
27    # Активационная функция
28    def sigmoid(self, x):
29        return 1 / (1 + np.exp(-self.clamp(x)))
30
31    # Метод предсказания нейронной сети
32    def predict(self, inputs): # inputs - массив с входными данными
33        inputs_1 = np.dot(self.weights_1,2, inputs) # Умножаем веса входного слоя на входные значения
34        outputs_1 = self.sigmoid_function(inputs_1) # Присваиваем к выходным значениям сигмоид
35
36        # Умножаем веса (скрытого слоя - выходного слоя на каждый элемент
37        inputs_2 = np.dot(self.weights_2,3, outputs_1)
38        outputs_2 = self.sigmoid_function(inputs_2) # Присваиваем к выходным значениям сигмоид
39        return outputs_2 # Результат предсказания
40
41    # Метод обучения по алгоритму обратного распространения сигнала, на вход передается массив входных значений и
42    # ожидаемый прогноз
43    def train(self, inputs, expected_predict):
44        # Анонимная функция predict
45        inputs_1 = np.dot(self.weights_1,2, inputs)
46        outputs_1 = self.sigmoid_function(inputs_1)
```

Пример работы в Python

Порог вхождения в языки программирования самый высокий по сравнению с другими инструментами, так как нужны специальные знания в области IT и статистики, а также умение писать код. Нельзя просто прочитать инструкцию для «чайников» и пойти программировать работающие системы. Ведь между копированием библиотеки и полноценным решением огромная разница.

Для бизнеса немаловажное значение имеет, как быстро и сколько сотрудников смогут разрабатывать решения на языке программирования. Сейчас много доступных обучающих курсов на популярных площадках, но:

- нужно длительное время на изучение;
- высок шанс для непрограммиста осознать, что это не его и он не сможет осилить эти знания.

Многие начинающие аналитики, хоть и не признаются в этом, действительно не сумели стать разработчиками на Python или R и за год обучения.

Резюме:

+ : Если предыдущие инструменты не решают ваших задач, то переходите на новый уровень прокачки своих аналитических умений — изучайте языки программирования. С помощью них вы сможете настроить весь процесс анализа данных и использовать, в том числе, продвинутые алгоритмы машинного обучения в своей работе.

— : Для начинающего аналитика этот порог входа самый высокий. Помимо знаний в Data Science необходимы умения в области программирования. Будьте готовы, что на довольно плотное обучение уйдёт минимум полгода. Ведь бизнес-пользователь должен освоить новую, достаточно сложную, специальность.

Аналитические low-code платформы

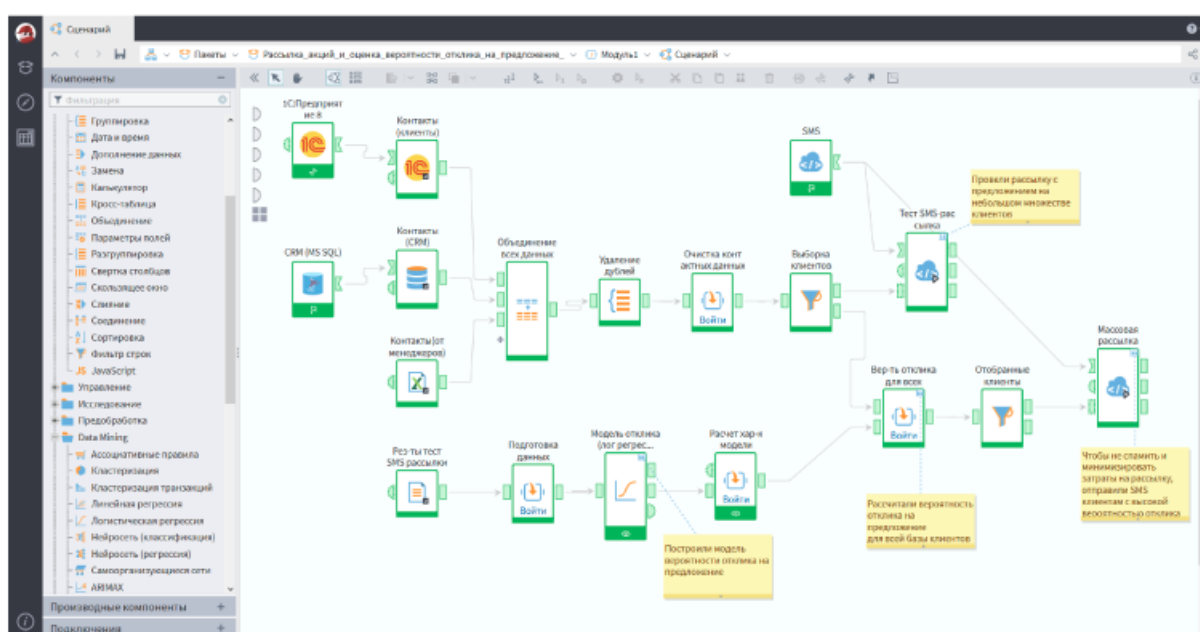
Здесь речь пойдёт не про все аналитические платформы, а только те, которые работают по принципу low-code. Эти инструменты визуального проектирования были разработаны специально для аналитиков, не обладающих навыками программирования, и оснащены

всеми необходимыми инструментами для простой работы с данными. Примеры таких решений: Loginom, Alteryx и т.д.

Аналитические платформы, которые базируются на принципе low-code, представляют собой конструкторы с набором готовых «кубиков». Решения, которые ранее разрабатывались программистами, теперь могут собираться самими аналитиками, «мышкой», в короткие сроки. Системы показывают высокую производительность при работе с большими массивами данных.

Платформы позволяют автоматизировать ежедневную работу аналитика различной сложности, практически не привлекая к ней разработчиков. Получение данных из различных систем, объединение, преобразование, очистка, простые и сложные вычисления, визуализация — та самая работа, на которую у аналитиков уходит до 80% времени. Она легко выполняется на аналитической платформе без кодирования и не требует специфичных знаний.

Для продвинутой аналитики платформы содержат инструменты Machine Learning. Наличие упрощённых мастеров настройки алгоритмов анализа данных с уточняющей документацией максимально упрощает вход в профессию аналитика.



Пример работы в Loginom

Для применения алгоритмов продвинутой аналитики всё-таки понадобится изучение теории по анализу данных и математической статистике. Не требуется становиться 100% Data Scientist'ом, но должно быть понимание, для чего нужен определённый алгоритм анализа данных, как правильно подготовить данные для него и интерпретировать результаты.

Минусом аналитических платформ также является ограниченное количество компонентов. При нехватке функционала придётся использовать встроенные языки программирования и просить помощи у своих IT-шников. Low-code не исключает написание кода, а сводит его к минимуму.

Резюме:

+ : Визуальное проектирование понятно всем, кто работает в Excel. Для получения первых результатов непрофессиональным разработчикам достаточно пары дней. На базе аналитической платформы начинающие аналитики смогут реализовать большую часть своих ежедневных задач: от подготовки данных до машинного обучения и моделирования.

— : Для использования продвинутой аналитики придётся погуглить про алгоритмы анализа данных, изучить, что это такое и как может быть применено к вашим данным. В случае выхода за рамки low-code идеологии требуется написание кода или помощь IT-отдела.

Инструмент для аналитика данных — какой в итоге выбрать?

Анализ данных расширяет возможности компании, позволяя бизнесу получать инсайты. Ключевая роль в этом процессе теперь отводится бизнес-экспертам как основным «носителям» знаний. Бизнес-пользователю нужен лишь подходящий инструмент.

Мы перечислили инструменты, которые обеспечивают лёгкий доступ к данным и аналитике для начинающего Citizen data scientist. Каждый из них используется в подходящей для него области. Выбирайте самый простой из возможных способов решения вашей задачи.

Если можно обойтись Excel, то используйте Excel. Если начинают создаваться обходные решения или неэффективно используются существующие инструменты, переходите на новый уровень. Каждый последующий шаг — это открытие новых горизонтов и профессиональное развитие в самом перспективном IT-направлении – Data Science.

Развивайте свои аналитические навыки, пробуйте и выбирайте подходящий для вас инструмент для работы с данными. Ищите то решение, которое быстро и эффективно реализует вашу задачу, а главное, поможет избежать ежедневных рутинных операций.

2.2. Анализ ABC (ABC-analysis)

Метод, позволяющий классифицировать ресурсы компании по степени их важности и прибыльности. В его основе лежит принцип Парето: 20% всех товаров дают 80% оборота.

По отношению к ABC-анализу принцип Парето может прозвучать так: надежное наблюдение за 20% позиций позволяет на 80% контролировать систему, будь то запасы сырья и комплектующих, продаваемые товары и т.д.

В процессе ABC-анализа товары делятся на три категории: **A** — наиболее ценные, **B** — промежуточные, и **C** — наименее ценные. По сути, ABC-анализ — это ранжирование ассортимента по различным параметрам. Классифицировать так можно и поставщиков, покупателей, складские запасы, т.е. все, что имеет достаточное количество статистических данных. Результатом является группировка объектов по степени влияния на общий результат.

ABC-анализ основан на принципе дисбаланса. При его проведении строится график зависимости совокупного эффекта от количества рассмотренных элементов. Его называют кривой Парето, кривой Лоренца или ABC-кривой.

По результатам анализа ассортиментные позиции ранжируются и группируются в зависимости от размера их вклада в совокупный эффект. В логистике ABC-анализ обычно применяют с целью отслеживания объемов отгрузки определенных товаров и частоты обращений к ним, а также для классификации клиентов по количеству или объему сделанных ими заказов.

Процедура проведения ABC-анализа содержит следующие шаги:

- определяется цель анализа;

- определяются действия по итогам анализа (что делать с полученными результатами);
- выбирается объект и параметр анализа (по какому признаку он будет проводиться);
- составляется рейтинговый список объектов по убыванию значения параметра;
- рассчитывается доля параметра от общей суммы параметров с накопительным итогом. Доля с накопительным итогом вычисляется путем прибавления параметра к сумме предыдущих параметров;
- выделяются группы А, В и С.

Как правило, объектами ABC-анализа являются товарные группы, товарные категории, отдельные товары, поставщики. Каждый из них имеет разные параметры описания и измерения: объем продаж (в денежном или количественном измерении), доход (в денежном измерении), товарный запас, оборачиваемость и т.д.

Выбор пороговых значений параметра, по которым производится отнесение объекта к одной из категорий, как и в XYZ-анализе, зависит от конкретных особенностей решаемой задачи. Например, можно предположить, что узкий ассортимент категории А из 10% товаров дает 70% дохода.

Из оставшихся товаров 20% (категория В) дают 20% дохода, а прочие 70% (категория С) — всего 10%. Из полученных результатов можно сделать вывод, что наибольшее внимание следует уделять товарам категории А, несколько меньшее — категории В, а товары категории С можно вообще рассматривать как вспомогательные.

2.3. Анализ XYZ (XYZ analysis)

Синонимы: XYZ-анализ

Разделы: [Бизнес-задачи](#)

Решения: [Loginom Customer Segmentation](#)

XYZ-анализ - метод классификации ресурсов компании в зависимости от:

- стабильности потребления и точности его будущего [прогнозирования](#);
- устойчивости или изменчивости спроса;
- подверженности сезонным колебаниям спроса;
- случайного характера спроса.

Алгоритм реализуется следующим образом:

- Определяем [коэффициенты вариации](#) для анализируемых ресурсов.

$$v = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}}{\bar{x}} \cdot 100\%,$$

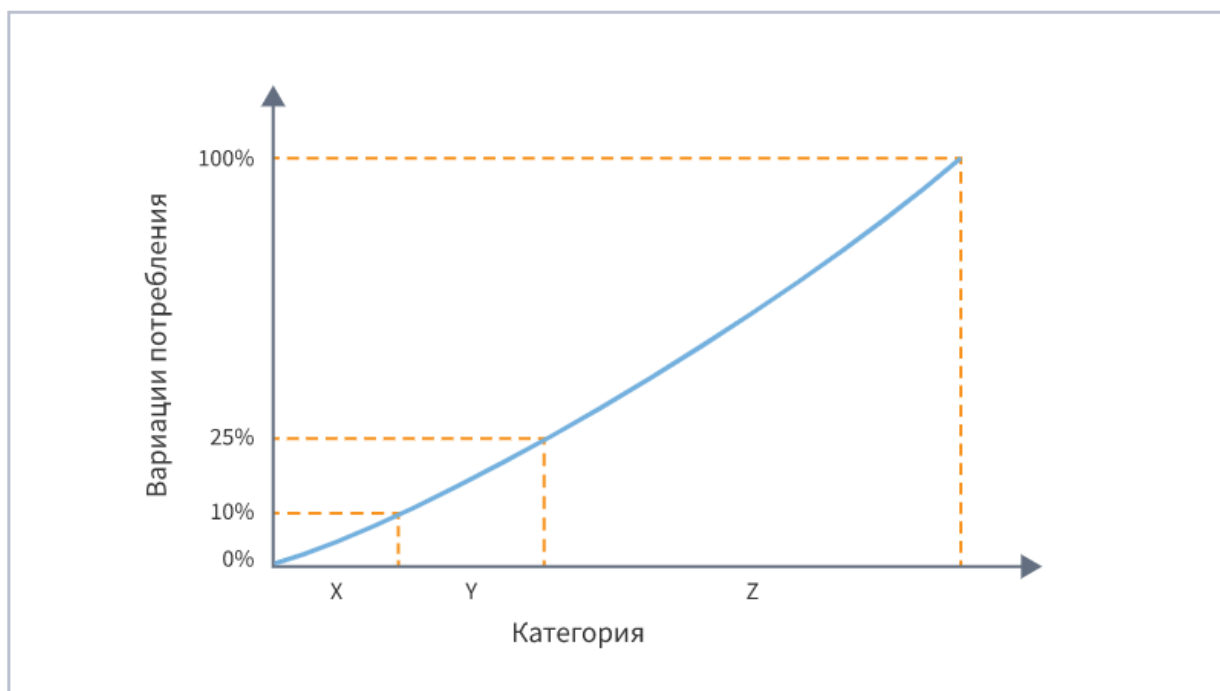
где x_i – значение параметра по оцениваемому объекту за i -тый период, \bar{x} - среднее значение параметра по всем периодам, n – число периодов.

- Сортируем ресурсы в соответствии с возрастанием коэффициента вариации.
- Распределяем ресурсы по категориям X, Y, Z:
 - **Категория X** — ресурсы характеризуются стабильной величиной потребления, незначительными колебаниями в их расходе и высокой точностью [прогноза](#). Значение коэффициента вариации находится в интервале от 0 до 10 %.
 - **Категория Y** — ресурсы характеризуются известными тенденциями определения потребности в них (например, [сезонными](#)

колебаниями) и средними возможностями их прогнозирования. Значение коэффициента вариации — от 10 до 25 %.

- **Категория Z** — потребление ресурсов нерегулярно, какие-либо тенденции отсутствуют, точность прогнозирования невысокая. Значение коэффициента вариации — свыше 25 %.

Графически интерпретируем результаты анализа.



XYZ-анализ представляет интерес для дистрибьюторов и производителей, имеющих свои склады. Он позволяет снижать издержки компании, связанные с логистикой и хранением ресурсов, а также с прямыми рисками (например, списанием товара по сроку годности).

2.4. ABC XYZ — анализ продаж для умного управления бизнесом

Проведение ABC XYZ-анализа стратегически важно для бизнеса. Такой подход позволяет получить информацию о продажах и ответить на вопросы:

- какие товары стоит запасать на склад;
- каким образом скорректировать ассортимент;
- кому из клиентов уделить больше внимания;
- кто из партнеров приносит прибыль, а кто убытки.

ABC-анализ

ABC подход позволяет классифицировать ресурсы компании по степени их важности. В его основе лежит принцип Парето: 20% усилий обеспечивают 80% результата.

В итоге, получается следующая градация категорий по степени влияния на общий результат:

- А — наиболее ценные. Это 20% от всей номенклатуры, которые приносят 80% прибыли.
- В — промежуточные. 30% позиций, на долю которых приходится 15% доходности.
- С — наименее ценные. Оставшиеся 50%, приносящие всего 5% от всех доходов.

На практике цифры могут отличаться.

Важно: новые продукты следует исключить из списка, так как они еще не успели себя зарекомендовать и могут стать кандидатами на вылет, хотя в них есть потенциал.

XYZ-анализ

XYZ подход применяется для определения характера спроса. С помощью него можно определить насколько будут стабильными продажи или услуги, дать анализ поведения клиентов в разное время.

Группировка происходит следующим образом:

- X — стабильный спрос, высокая точность прогноза.

Значение коэффициента вариации в пределах от 0 до 10%.

- Y — нерегулярные заказы, есть сложности с прогнозированием. Могут зависеть от сезонности. Значение коэффициента вариации от 10% до 25%.

- Z — нерегулярное потребление, единичные заказы, точность прогнозирования низкая. Коэффициент вариации – от 25%.

Важно: следует учитывать сезонность, акции, тренды и другие краткосрочные тенденции.

ABC XYZ-анализ

ABC XYZ-анализ продаж — это отдельные независимые методики. Но наибольшую эффективность показывает их совместное использование и наложение результатов друг на друга. Таким образом, создается эффект синергии, и получается более полная картина, чем при их раздельном применении.

ABC-анализ помогает определить товары, которые приносят максимальную прибыль, а XYZ-анализ дает информацию об их стабильности.

Почему не стоит использовать Excel для ABC XYZ-анализа

Часто произвести расчеты ABC XYZ анализа предлагают в Excel. Действительно, функционал электронных таблиц позволяет производить вычисления. Но, когда речь идет о реальном бизнесе с огромной номенклатурой, множеством точек продаж, развитой партнерской сетью, — использование Excel только затрудняет бизнес-процессы.

Проблемы начинаются сразу же:

- Представьте, что вам нужно сделать ABC XYZ-анализ для хотя бы 50000 товарных позиций. Уверены, что хотите сделать это вручную?
- При большом количестве данных, даже на открытие файла потребуется затратить время. Что уж говорить про вычисления.
- Excel позволяет использовать только 1 048 576 строк. Если данных больше, то уже не важно, насколько быстро он будет работать.
- Проблема с открытием больших файлов — довольно распространенное явление в работе Excel. Связана она с использованием вычислительных функций, условным форматированием, созданием больших массивов данных и сводных таблиц и др.
- Повторное использование и формирование отчетов затруднено из-за «тормозов» и больших массивов данных. После каждого действия приходится ждать.

Перечисленных выше пунктов достаточно, чтобы сделать вывод: для реальных вычислений нужен более продвинутый инструмент.

ABC XYZ-анализ в Loginom

Рассмотрим реальную бизнес задачу на примере организации, поставляющей товары для медицинских учреждений.

Цель анализа: сегментация товаров для управления складскими запасами.

Для достижения поставленной цели проводим ABC XYZ-анализ с помощью Loginom.

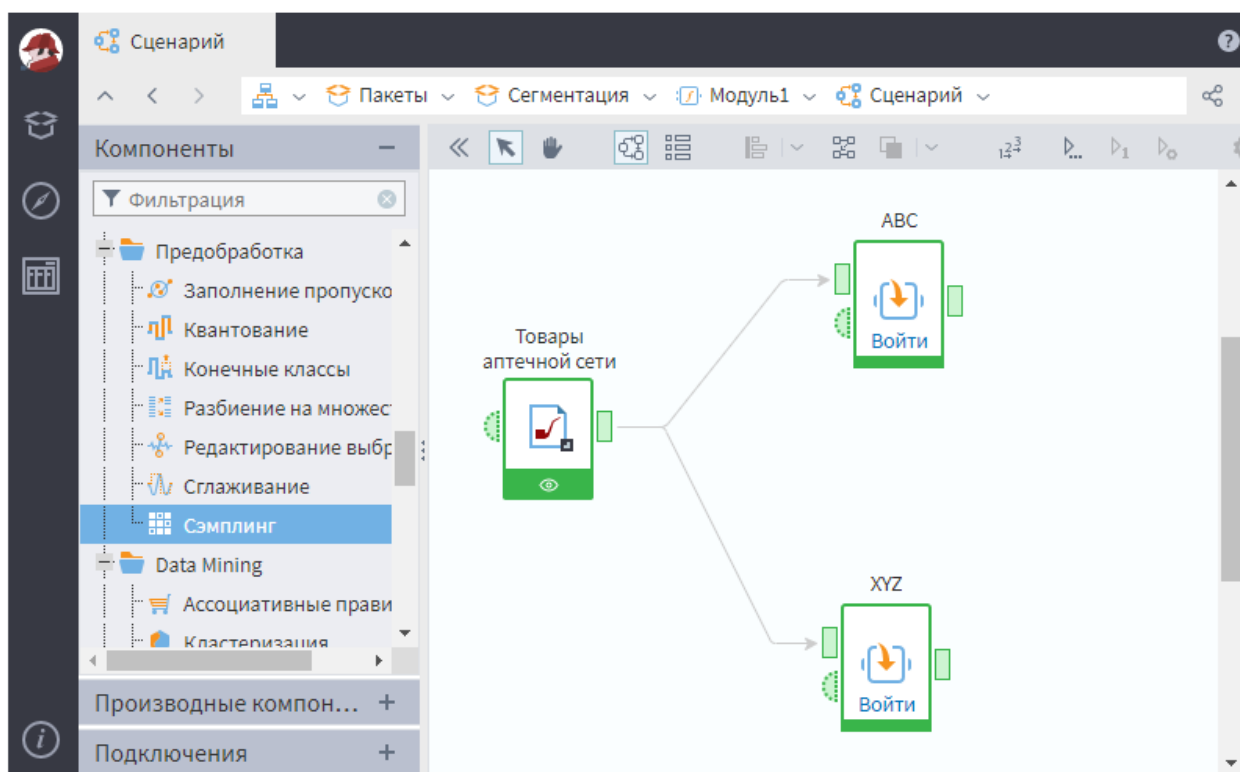
На первом этапе выгружаем данные по всем товарам за всё время работы. Это более 1 миллиона записей.

#	12 Н...	ab Медикамент	9.0 Ц...	9.0 Су...	9.0 К.	31 Дата в...	31 Дата ...	ab f
47	324	Салфетки влажные NUANCE ан...	19.93	99.65	5	7/7/2003	7/7/2003	
48	324	Ватные палочки 100шт. (полиэ...	8.56	85.6	10	7/7/2003	7/7/2003	
49	54	Саргиссio Крем депилят.д/лица...	93.12	186.24	2	7/2/2003	7/2/2003	
50	54	БАД Диетический коктейль Кл...	320.74	320.74	1	7/2/2003	7/2/2003	
51	54	Крем-лосьон СМИЛЬД фл. 50мл	29.9	59.8	2	7/2/2003	7/2/2003	
52	6610	Панкреатин таб. 25ЕД х60	19.56	195.6	10	11/4/2003	11/4/2003	
53	6610	Антигриппин д/взросл. шип. т...	19.03	570.9	30	11/4/2003	11/4/2003	
54	5923	Эм-эвал Детский блистер 31г	16.67	483.43	29	10/22/2003	10/22/2003	
55	5923	Шприц Микрофайн 1мл инсул...	2.56	768	300	10/22/2003	10/22/2003	
56	5923	Нистатин таб. 500тыс.ЕД х20	10.54	316.2	30	10/22/2003	10/22/2003	
57	5923	Сульгин таб. 0.5г х10	1.65	82.5	50	10/22/2003	10/22/2003	
58	5923	Шприц одноразовый 5мл ВЕСТ...	1.33	665	500	10/22/2003	10/22/2003	
59	5923	Шприц одноразовый 10мл ВЕС...	1.73	519	300	10/22/2003	10/22/2003	

Исходный набор данных о продажах

Процесс загрузки занял 1 секунду.

Для каждого типа анализа построим свой сценарий работы.



Сценарий ABC XYZ в Logint

Результаты вычислений сводим в одну таблицу.

#	ab Объект	ab ABC XYZ	ab ABC категория	ab XYZ категория
109	Перчатки смотровые латексные н/ст опудре...	AZ	A	Z
110	Реамберин 1,5% фл. 400мл	AZ	A	Z
111	Новокаин амп. 0,5% 5мл x10	AZ	A	Z
112	Пленка RETINA XBM 30x40 (100л)	AZ	A	Z
113	Шприц одноразовый 20мл HELM (Германия)	AZ	A	Z
114	Сульфаниламид (пара-аминобензолсульфам...	BX	B	X
115	Полипропилен, кг	BX	B	X
116	Гентамицина сульфат стерильный	BX	B	X
117	Цефотаксим (субстанция)	BX	B	X
118	Катетер для периф.вен 22G	BZ	B	Z
119	Салфетки д/сервировки PALOMA VELIKONOC...	BZ	B	Z
120	Форамен Детский набор: з/щ мягкая +песочн...	BZ	B	Z
121	Шефилд з/п WHITENING с отбеливающим э...	BZ	B	Z
13 670	Шприц одн. 1мл. тубер. (игла импортная)	BZ	B	Z

Результат

Таблица: интерпретация совмещенных результатов

AХ	AУ	AZ
Топовая категория. В нее попадают объекты со стабильно высокой доходностью и прогнозируемым спросом.	Здесь также заложена хорошая прибыль, но продажи не такие предсказуемые.	Хороший заработок, но кривая имеет значительные отклонения. Другими словами: «то густо, то пусто».
BХ	BУ	BZ
Средние, но стабильные доходы. Легко смоделировать дальнейшие действия.	Нестабильное потребление при усредненном показателе маржинальности.	Непредсказуемый спрос в совокупности со средними значениями по выручке.
CХ	CУ	CZ

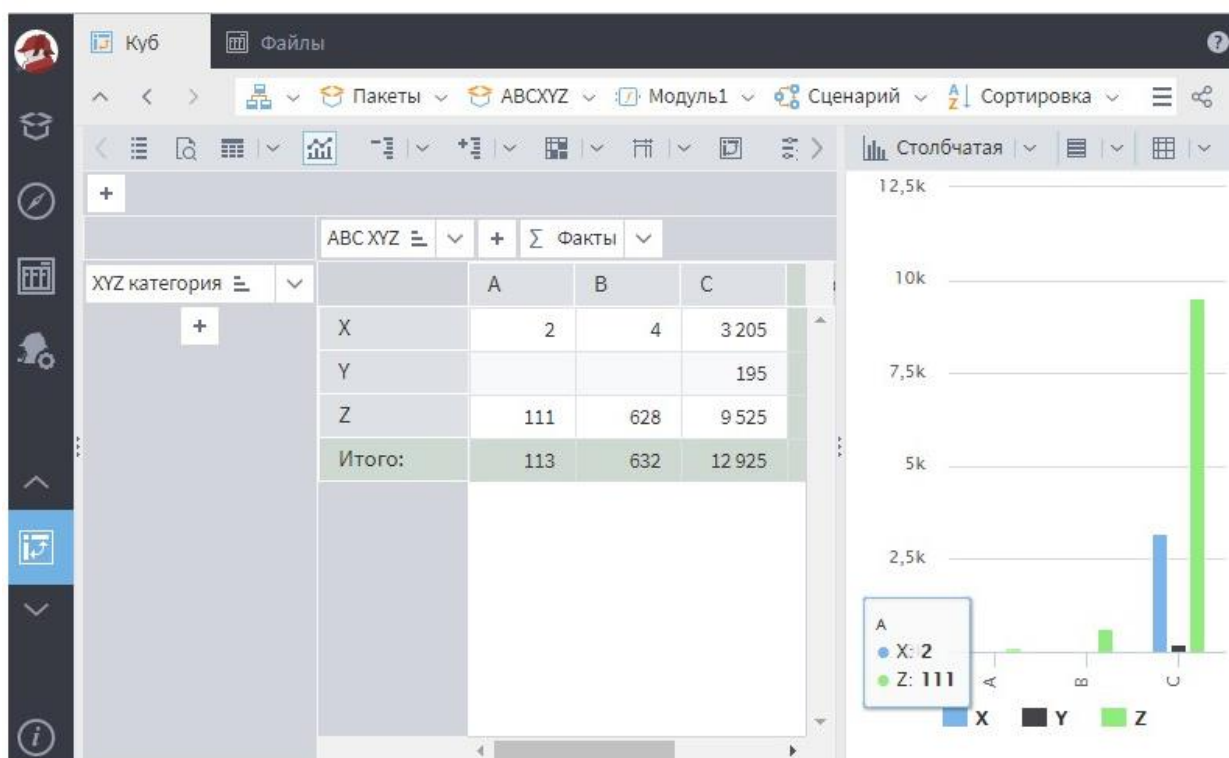
AX	AY	AZ
Низкая стоимость со стабильным прогнозом. «Мал золотник, да дорог».	Маленькая цена, и нерегулярный спрос.	Ни заработка, ни постоянства. От таких объектов в 99% случаев нужно избавляться.

Визуализация результатов ABC XYZ-анализа

Logiном позволяет удобно визуализировать данные.

Построим OLAP-куб для наглядного представления результатов.

Получилась следующая таблица:



OLAP-куб

Как видно из нашего примера, товаров типа AX в компании всего 2. Товаров типа AY, которые также приносят высокий стабильный доход, нет совсем. Основная масса — это небольшие нерегулярные продажи.

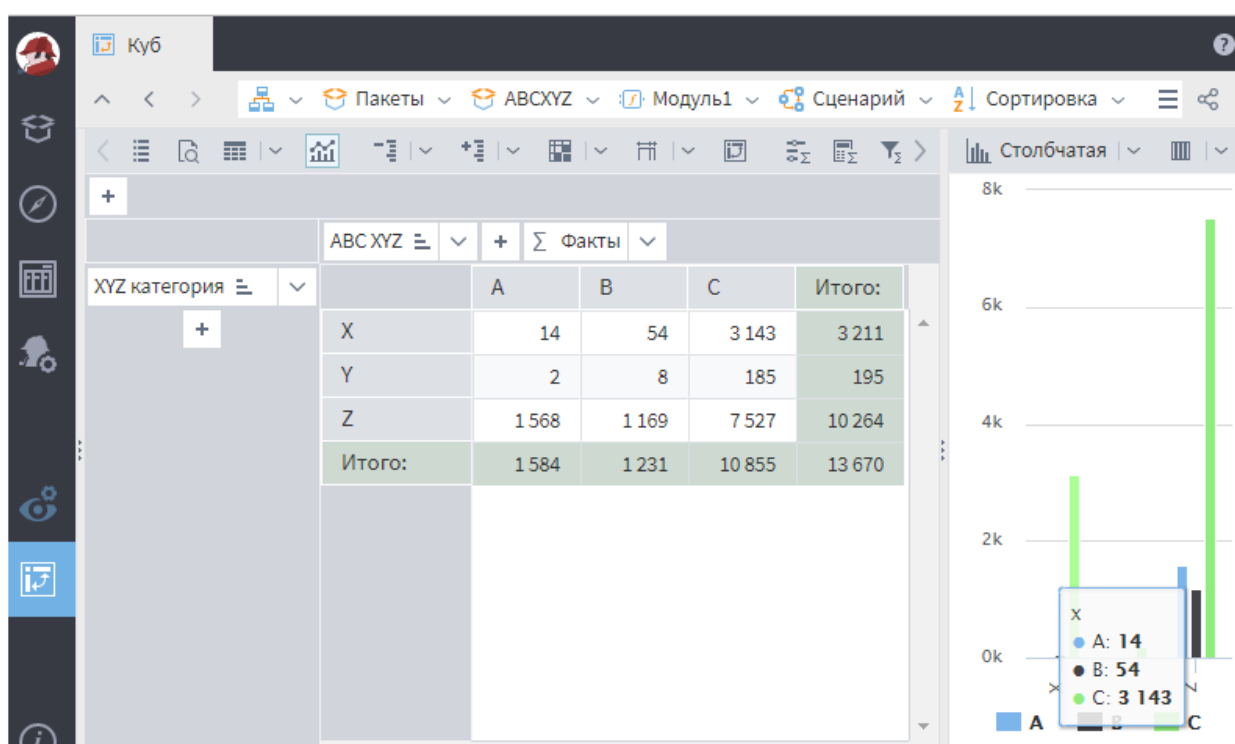
Неудачно подобранные границы могут привести к тому, что почти все товары попадут в один сегмент. Это лишает смысла применение ABC XYZ-анализа, т.к. целью является выстраивание работы с каждой

группой по-разному. В случае, когда в итоговой таблице оказываются пустые ячейки или данные распределяются слишком неравномерно, следует пересмотреть границы ABC и XYZ.

Эмпирическим путем изменим границы ABC-анализа следующим образом:

- А — нижняя граница 90%.
- В — нижняя граница 95%.
- С — все оставшиеся.

Пустых ячеек больше нет, результат выглядит следующим образом:



OLAP-куб

Для проведения ABC-анализа не обязательно руководствоваться принципом Парето. Границы можно менять до тех пор, пока данные в итоговой таблице не будут оптимальными для дальнейшей работы.

Отчеты ABC XYZ-анализа являются источником важной информации. В разрезе большого промежутка времени намного

эффективнее сконцентрироваться на тех товарах, которые приносят максимальную прибыль.

Кроме того, необходимо наблюдать за поведением отчета в динамике. В разные периоды товары могут изменять категории с одной на другую. То, что сейчас товар находится в категории АХ, не значит, что в следующем квартале он будет там же. ABC XYZ-анализ выявит, какой товар приносил прибыль и когда перестал. Это позволит эффективно управлять складскими запасами и быстро принять меры по восстановлению спроса.

Loginom — решение для бизнеса

Loginom – это надежный, быстрый инструмент для работы с большими массивами данных:

- Позволяет быстро, без «тормозов» загружать любые объемы.
- Выполняет сложные вычисления «на лету».
- Сценарный принцип работы без привязки к конкретным данным позволяет создавать подмодели и использовать в других сценариях.
- Визуализация отчетов позволяет наблюдать за изменениями и перетоками данных в динамике в различных разрезах.

2.5. Анализ RFM (RFM-analysis)

RFM-анализ - это анализ клиентов компании с целью их сегментации по ценности для бизнеса. Широко используется в маркетинге, директ-маркетинге и, особенно, в розничной торговле и сфере услуг.

Анализ использует 3 измерения:

- **Recency (давность)** — как давно клиент совершил последнюю покупку? Чем меньше времени прошло с момента последней активности клиента, тем больше вероятность, что он повторит действие.
- **Frequency (частота)** — как часто клиент совершал покупки? Чем больше покупок совершит клиент, тем больше вероятность того, что он их повторит в будущем.
- **Monetary (деньги)** — сколько клиент тратит? Чем больше денег было потрачено, тем больше вероятность того, что клиент будет совершать покупки и в дальнейшем.

Предполагается, что клиент, проявивший себя недавно, показывающий повышенную активность и тратящий на покупки больше денег, будет проявлять активное потребительское поведение и в дальнейшем.

История покупок клиента или продаж товара представляется в виде таблицы или RFM-матрицы, в которой три колонки – давность, частота и деньги. Каждая колонка разбивается на категории.

Например, давность можно разбить на интервалы 1 - 30 дней (настоящее время), 31 – 60 дней (недавно), 61 – 90 дней (давно). Частоту покупок можно разделить на частые (более 10 в месяц), редкие (3 - 10 в месяц) и разовые (менее 3-х раз в месяц). Затраты

клиентов можно разделить на высокие (5 - 10 тыс.), средние (2 - 5 тыс.) и низкие (менее 2 тыс.). Тогда можно составить RFM-матрицу вида:

Частота	Давность	Деньги		
		Высокие	Средние	Низкие
Часто	Настоящее время			
	Недавно			
	Давно			
Редко	Настоящее время			
	Недавно			
	Давно			
Разово	Настоящее время			
	Недавно			
	Давно			

Нетрудно увидеть, что наиболее «перспективные» клиенты окажутся в верхнем левом углу таблицы, в выделенном сегменте (содержит клиентов, которые делают покупки часто, в настоящее время или недавно, тратя при этом средние или большие суммы). Наименее «перспективные» окажутся в нижнем выделенном сегменте (разовые покупки, сделанные давно или недавно на низкие и средние суммы).

В зависимости от особенностей анализа могут использоваться и другие представления RFM-матрицы. Например, иногда используют не 3 категории для каждого измерения, когда получается 27 сегментов, а 5 категорий, что позволяет создать 125 сегментов. В этом случае результаты анализа оказывается более детальными.

Иногда сегментам матрицы присваивают баллы от 1 (самые «непривлекательные») до 5 (самый «привлекательный» сегмент). В этом случае «лучший» сегмент будет обозначен 5R-5F-5M, а «худший» 1R-1F-1M. Тогда клиентов, попавших в сегмент 3R-3F-3M можно интерпретировать как средних по привлекательности.

Преимуществом метода является простота (не требует специального статистического ПО), а результаты легко интерпретируются. При использовании в директ-маркетинге применение анализа RFM может увеличить количество откликов на рекламные акции.

Существуют несколько модификаций RFM-анализа:

- RFD – Recency (давность), Frequency (частота), Duration (продолжительность) - модифицированная версия RFM-анализа для изучения поведения потребителей бизнес-продуктов, ориентированных на зрителей, читателей, интернет-сёрферов (например, количество времени, проведенного серферами в онлайн-кинотеатре).
- RFE - Recency (давность), Frequency (частота), Engagement (вовлечённость) расширенная версия RFD-анализа, где можно оценить вовлечение клиента в работу с бизнес-продуктом. Кроме продолжительности при этом используют, например, количество посещённых web-страниц, количество переходов по ссылкам и другие действия клиента, свидетельствующие о его активном использовании бизнес-продуктом. Предполагается, что наиболее активные клиенты более склонны к отклику на маркетинговые акции.
- RFM-I – Recency (давность), Frequency (частота), Monetary – Interactions (деньги-взаимодействие) - является версией RFM-анализа для учета давности и частоты маркетинговых взаимодействий с клиентом, например, для изучения возможных сдерживающих эффектов частых рекламных мероприятий (когда слишком назойливая реклама отталкивает клиента).