Департамент образования и науки города Москвы

Государственное автономное образовательное учреждение высшего образования города Москвы

«Московский городской педагогический университет»

Институт цифрового образования

Департамент информатики, управления и технологий

Кузьменко Сергей Викторович

**КУРСОВАЯ РАБОТА**

Тема: «Современные подходы для построения облачной продуктовой аналитики»

Направление подготовки

38.04.05 Бизнес-информатика

Профиль образовательной программы

Бизнес-аналитика и большие данные

Курс обучения: 1 курс магистратуры

Форма обучения: очная

Руководитель: Сахнюк П.А., к.т.н., доцент

Москва

2024

**СОДЕРЖАНИЕ**

[**ВВЕДЕНИЕ** 3](#_Toc162517688)

[**ГЛАВА 1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ** 4](#_Toc162517689)

[**1.1.** **Применение облачных технологий в продуктовой аналитике** 4](#_Toc162517690)

[**1.3. Методика проведения исследовательского анализа данных** 5](#_Toc162517691)

[**ГЛАВА 2. ПРАКТИЧЕСКАЯ ЧАСТЬ** 8](#_Toc162517692)

[**2.1. Подготовка и анализ данных в Google Colab** 8](#_Toc162517693)

[**2.2. Анализ данных в Google BigQuery** 13](#_Toc162517694)

[**2.3. Загрузка и анализ данных в Loginom** 15](#_Toc162517695)

[**2.4. Кластеризация клиентов с применением машинного обучения в BigQuery.** 20](#_Toc162517696)

[**ЗАКЛЮЧЕНИЕ** 23](#_Toc162517697)

[**СПИСОК ЛИТЕРАТУРЫ** 24](#_Toc162517698)

# **ВВЕДЕНИЕ**

Облачная продуктовая аналитика стала неотъемлемой частью успешного ведения бизнеса в современном мире. Она позволяет компаниям анализировать данные о продуктах, клиентах и рынке для принятия обоснованных решений и повышения эффективности бизнес-процессов. Облачные сервисы имеют ряд преимуществ [1], связанных с масштабируемостью и гибкостью, что привело к росту популярности в последние несколько лет [2]. Этим обуславливается **актуальность** данной работы

**Целью** курсовой работы является построение аналитики исследуемого сервиса (продукта) с применением облачных решений Google Colab и Google BigQuery, а также с использованием локального ПО Loginom, обращающегося к облачному хранилищу, как к источнику данных.

Для достижения цели работы ставятся следующие **задачи**:

1. Изучение теоретических основ применения облачных технологий в аналитике продукта;
2. Проведение подготовки данных в Google Сolab и загрузка обработанных данных в облачный сервис Google BigQuery;
3. Подключение с локального устройства из программы для анализа данных Loginom к облачному хранилищу Google BigQuery;
4. Проведение ABC, XYZ и RFM анализа на всех трех платформах и сопоставление результатов;
5. Применение алгоритма K-means для кластеризации клиентов.

# **ГЛАВА 1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ**

# **Применение облачных технологий в продуктовой аналитике**

Применение облачных технологии подразумевают удаленное использование средств обработки информации и играют важную роль в современной продуктовой аналитике.

Рассмотрим основные применения подобных сервисов. Одним из ключевых аспектов является возможность хранения и обработки больших объемов данных, что особенно важно для продуктовой аналитики, где необходим доступ к большому количеству информации о продукте и пользователях. Также облачные технологии позволяют масштабировать инфраструктуру в зависимости от потребностей. В продуктовой аналитике объемы данных и запросов могут сильно различаться в зависимости от текущих задач. Облачные платформы обычно предлагают широкий набор инструментов для интеграции данных из различных источников, что позволяет продуктовым аналитикам работать с данными из различных систем и приложений. Облачные сервисы позволяют анализировать данные и строить отчёты в реальном времени, что дает для аналитиков возможность быстрой реакции на изменения в поведении пользователей. Облачные платформы предоставляют мощные инструменты для машинного обучения и аналитики, что позволяет улучшить процессы продуктовой аналитики и предсказывать поведение пользователей и основные продуктовые показатели.

Отметим основные преимущества облачных платформ [3] перед локальными сервисами:

1. Доступность. Такие сервисы доступны в любой точке мира. Необходим только интернет и браузер.
2. Низкая стоимость и простота. Облачные сервисы позволяют снижать затраты на поддержание инфраструктуры. Всю поддержку и ответственность за обеспечение работоспособности сервисов берет на себя компания, которая предоставляет ресурс. Клиент только платит за использование ресурса, получая готовое решение проблем бизнеса
3. Гибкость и масштабируемость. «Облако» дает практически неограниченные ресурсы по памяти и процессору. Все вычислительные затраты берет за себя сервис, что означает снижение технических требований для локального пользователя.
4. Безопасность. Компании, предоставляющие облачные сервисы обеспечивают защиту информации.

Таким образом, облачные технологии играют ключевую роль в развитии продуктовой аналитики, обеспечивая доступ к нужной информации, её обработку и анализ в реальном времени, что помогает развивать продукт и улучшать его пользовательский опыт.

# **1.3. Методика проведения исследовательского анализа данных**

Для демонстрации методов продуктовой аналитики с применением облачных технологий рассмотрим датасет “E-commerce data”, который содержит транзакции e-commerce сервиса с 2010 по 2011 год. Витрина данных представлена в виде файла формата CSV.

Информация об атрибутах набора данных:

1. InvoiceNo: Номер счета
2. StockCode: Код продукта
3. Description: Название продукта.
4. Quantity: Количество каждого продукта в рамках одной сделки (транзакции)
5. InvoiceDate: Дата и время транзакции (сделки).
6. UnitPrice: Цена за единицу продукта в фунтах стерлингов.
7. CustomerID: Номер клиента. Уникальный для каждого клиента.
8. Country: Название страны, в которой проживает клиент.

Для анализа данных применим ABC, XYZ и RFM анализы. ABC и XYZ можно применять как в разрезе по клиенту, так и по продукту. Последнее наиболее подходит к тематике работы. То есть задача состоит в том, чтобы понять, насколько хорошо продается тот или иной товар и в итоге каждому StockCode сопоставить определенный класс для каждого из видов анализа. RFM анализ применяется для сегментации клиентов.

Рассмотрим подробнее все три вида анализа.

ABC анализ [4] основан на принципе Парето «соотношение 80 на 20». В нашем случае с акцентом на продукт: 20 % товаров обеспечивают 80 % прибыли магазина. Поэтому для проведения анализа для необходимо сгруппировать данные по коду товара, вычислить процент продаж товара от общей суммы продаж, отсортировать по убываю и рассчитать кумулятивную сумму. Далее товары делятся на три класса сверху вниз:

1. Класс А: Первые товары, накопительная сумма которых меньше 20 %. Это товары, которые вносят основной вклад в формирование выручки
2. Класс B: Следующая группа товаров с накопительной суммой от 20 до 80 %. Это товары “среднего класса”. Влияют на формирование выручки, но не вносят определяющий вклад
3. Класс С: Оставшиеся товары с накопительной суммой от 80 до 100 %. Товары, которые меньше всего влияют на выручку магазина.

XYZ анализ [5] необходим для понимания стабильности спроса на товар. Для проведения такого анализа необходимо сгруппировать по товару и для каждого товара посчитать коэффициент вариации, как отношение стандартного отклонения количества единиц товара к среднему значению количества товара в одной транзакции для данного продукта . Затем товары разбиваются на три класса:

1. Класс X: Коэффициент вариации от 0 до 10 %. Данные продукты имеют очень низкую дисперсию, то есть потребляются стабильно. Для них легко рассчитать объемы потребления и предсказать выручку к концу периода
2. Класс Y: Коэффициент вариации от 10 до 25 %. Такие товары имеют больший разброс по количеству потребления, что означает сложность в прогнозировании. Необходимо изучать природу эти колебаний
3. Класс Z: Коэффициент вариации более 25%. Говорит о случайности спроса. Такие товара сокращаются до минимума или исключаются из ассортимента.

RFM-анализ [6] применяется для сегментации клиенты, чтобы лучше понимать пользовательский опыт. Рассматриваются три показателя. Recency (давность) – количество дней с последней сделки. Frequency – частота. Как часто клиент совершает сделки. Monetary – заработанные деньги с клиента. Далее по каждому показателю данные сортируются (для Recency в порядке убывания, для Frequency и Monetary в порядке возрастания) и разбиваются по квантилям на заданное число групп.

Алгоритм кластеризации K-means входит в число базовых классических методов машинного обучения. Кластеризацию называют **обучением без учителя**. В пространстве признаков выбираются K центров. Каждый объект относится к тому классу, к центру которого он ближе. Далее на каждом шаге центры кластеров пересчитываются как центр масс всех объектов в этой группе. И заново происходит распределение объектов по кластерам. Продолжается это, пока центры кластеров не перестанут меняться.

В рамках данной курсовой работы приведенные выше исследования будут проведены в облачный платформах Сolab и BigQuery. А также в локальной программе Loginom [7], [8], использующего BigQuery как источник и облачное хранилище [9], [10].

# **ГЛАВА 2. ПРАКТИЧЕСКАЯ ЧАСТЬ**

# **2.1. Подготовка и анализ данных в Google Colab**

Для предварительной обработки была выбрана платформа Google Colab. Предобработка осуществляется на языке программирования Python. Набор данных в формате CSV выгружен с Google Disk и прочитан в формат Pandas DataFrame. На рисунке 1 показаны процессы загрузки и исходный датасет.



Рисунок 1 - Загрузка данных, импортирование библиотек и предварительный просмотр исходных данных

Следующим этапом были удалены строки с пропущенным ID клиента и преобразованы типы данных столбцов InvoiceDate и CustomerID, а также удалены дубликаты.

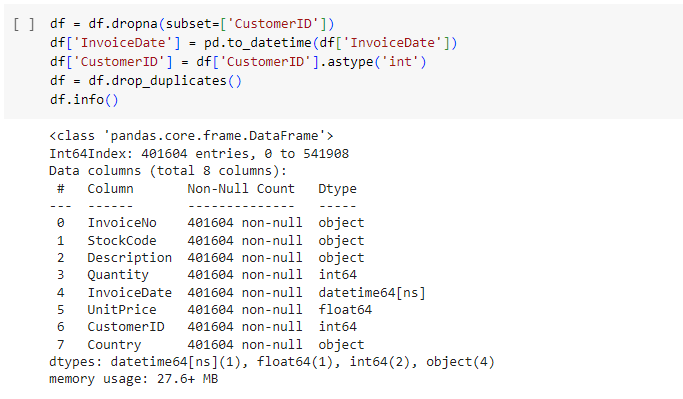


Рисунок 2 - Удаление пропущенных значений, дубликатов и преобразование типов

Затем было предрасчитано поле Costs (сумма по сделке) как произведение количества единиц товара Quantity и цены UnitPrice, и исключены строки с отрицательным Costs, так как это возврат товара, который не должен учитываться в анализе.

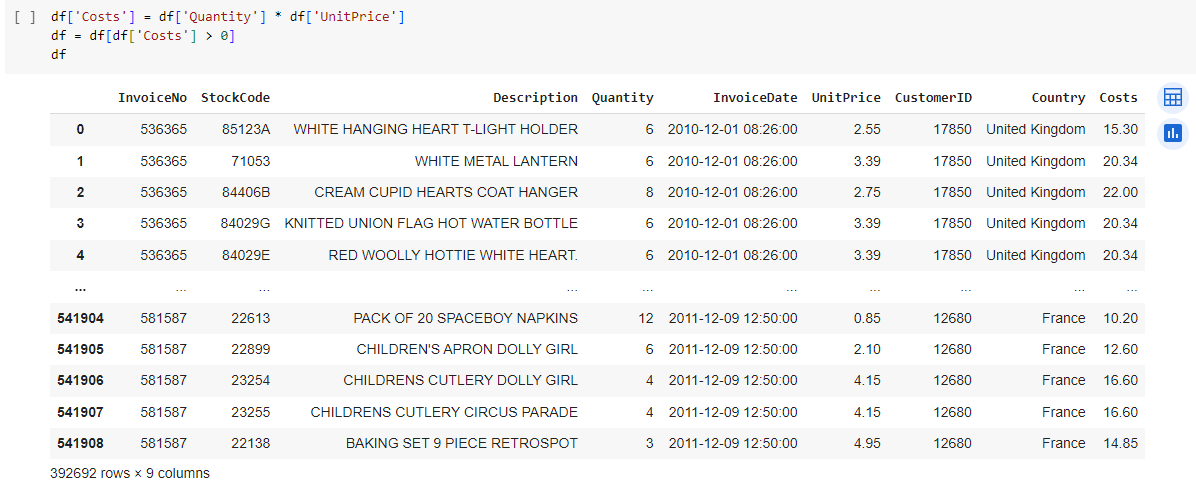


Рисунок 3 - Расчет Costs и исключение возвратов

Обработанный датасет был загружен в облачное хранилище BigQuery. Это требуется для дальнейшего анализа данных на платформе BigQuery и в Loginom.

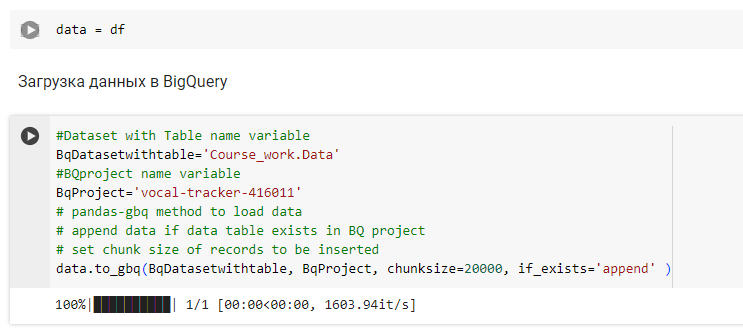


Рисунок 4 - Загрузка в BigQuery

Затем был проведен ABC анализ. Для этого данные были сгруппированы по коду товара StockCode, с рассчетом суммы по полю Costs. Далее был рассчитан процент от общей выручки (поле percent). После чего данные отсортировывались в порядке убывания по полю Percent и была высчитана накопительная сумма по этому полю. Товары с накопительной суммой до 20% процентов относятся с классу А, от 20 до 80 % к классу B, 80 – 100 % к классу С. После этого результат сегментации был добавлен с исходную таблицу (см. рисунок 5)

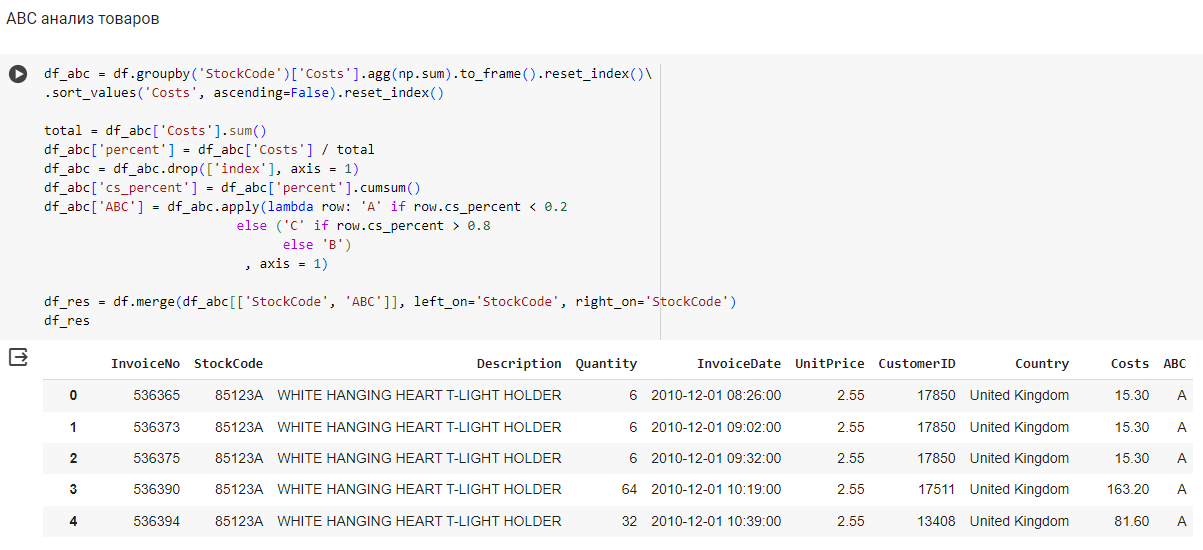


Рисунок 5 - Результат ABC анализа в Colab

Следующим этапом был проведен XYZ анализ. Для этого данные были сгруппированы по коду товара и рассчитан коэффициент вариации (см. Рисунок 6). Если коэффициент вариации равен 0 или больше 25 %, товар относится в классу Z. Коэффициент вариации от 0 до 10 % - класс X, от 10 до 25 % к классу Y. Результат присоединен к общему датасету.



Рисунок 6 - XYZ анализ к Colab

Третьим этапом проведен RFM анализ. Для это данные были сгруппированы уже по полю CustomerID, то есть данных вид анализа был произведен по клиентам. Показатель Recency рассчитан, как количество дней со дня последней покупки клиента до максимальной даты по всему датасету (2011-12-09 12:50:00). Frequency – количество уникальных заказов для каждого клиента, Monetary – общая выручка по клиенту (суммирование по полю Costs). Далее показатели были разделены по квантилям на 5 частей (причем показатель Recency был отсортирован по убыванию, а остальные 2 показателя по возрастанию). Результат добавлен к датафрейму df\_res, содержащего все предыдущие сегментации (см. Рисунки 7 - 8).



Рисунок 7 - Расчет Recency, Frequency, Monetary

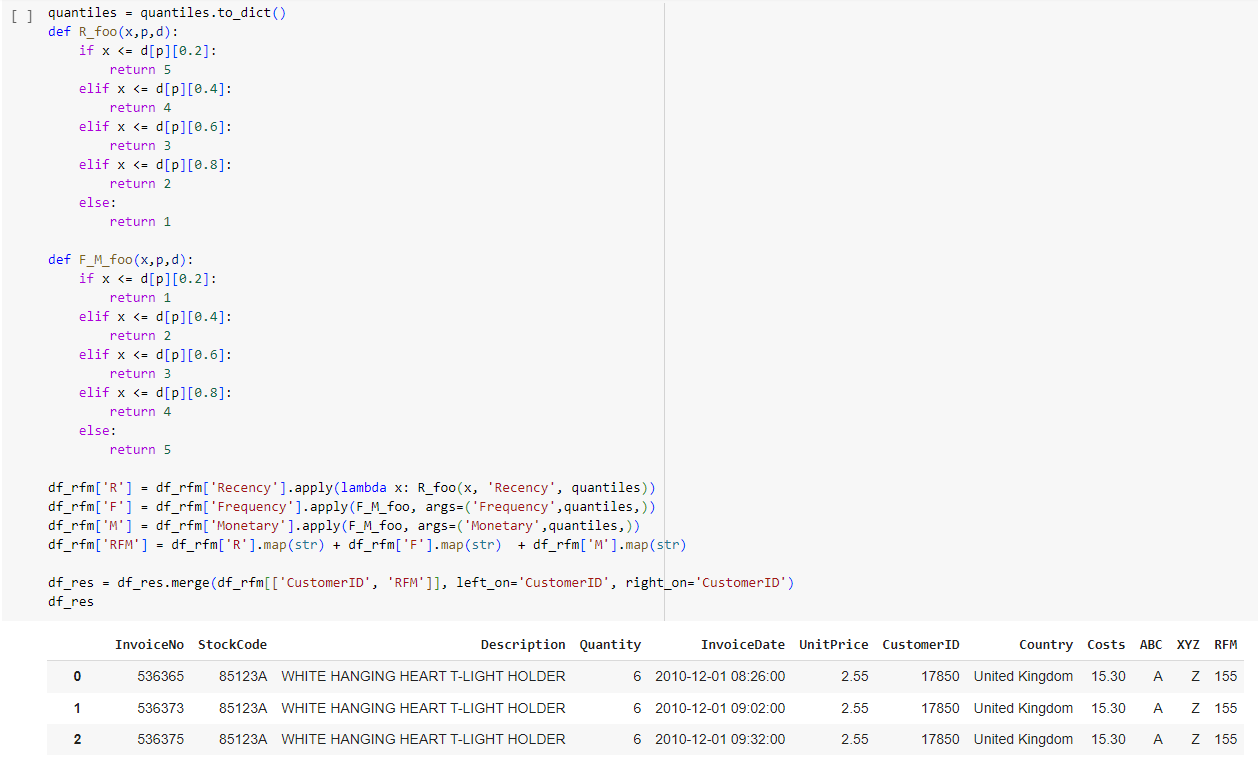


Рисунок 8 - Сегментация для RFM анализа

Таким образом были выполнены 3 вида анализа. ABC, XYZ по товарам и RFM для клиентов. Вместе с эти данные были загружены в облачное хранилище BigQuery для дальнейшей работы.

# **2.2. Анализ данных в Google BigQuery**

Загруженные в облачное хранилище данные можно анализировать при помощи SQL-запросов в BigQuery. Аналогично были рассчитаны все три вида анализа (см. Рисунки 9 - 11). Причем каждый анализ сохранен как представление (View). Преимуществом работы с представлениями вытекает из того факта, что предсатвление является агрегацией, которая используется в дальнейшем без лишних подзапросов. При этом само представление не является физической таблицей и не занимает места на диске. Представление обсчитывается при вызове в запросе. Использование сохраненных представлений для расчета итоговой таблицы с сегментациями товара и клиентов представлена на рисунке 12.

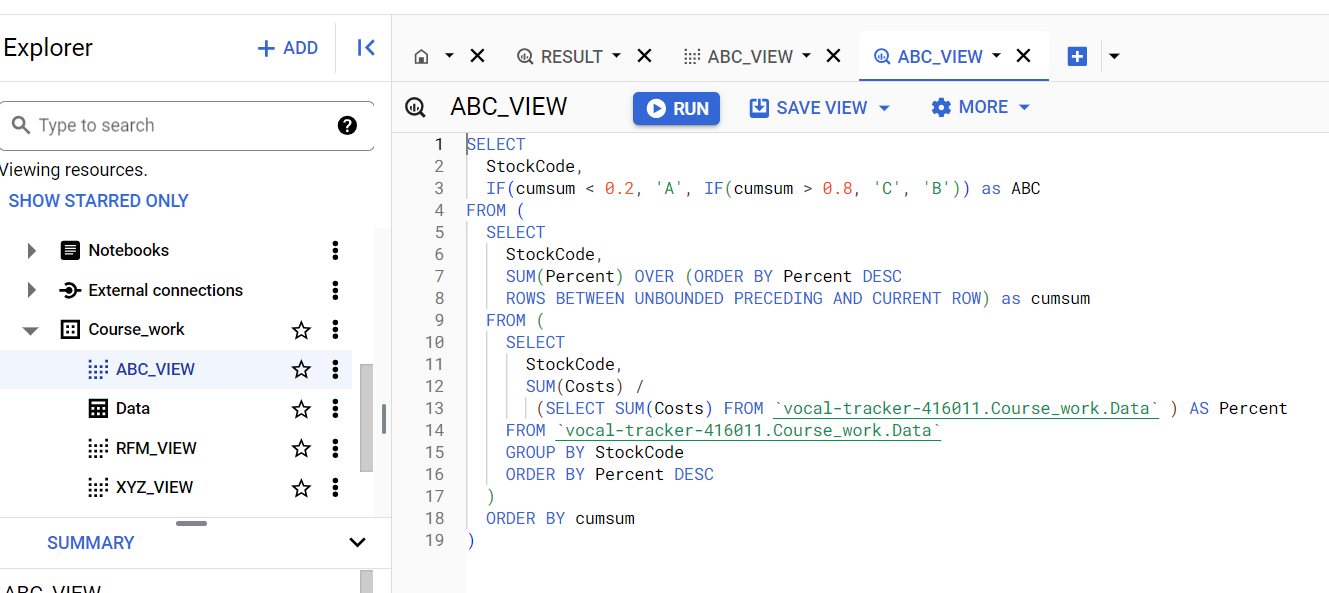


Рисунок 9 - ABC анализ в виде представления VIEW в BigQuery

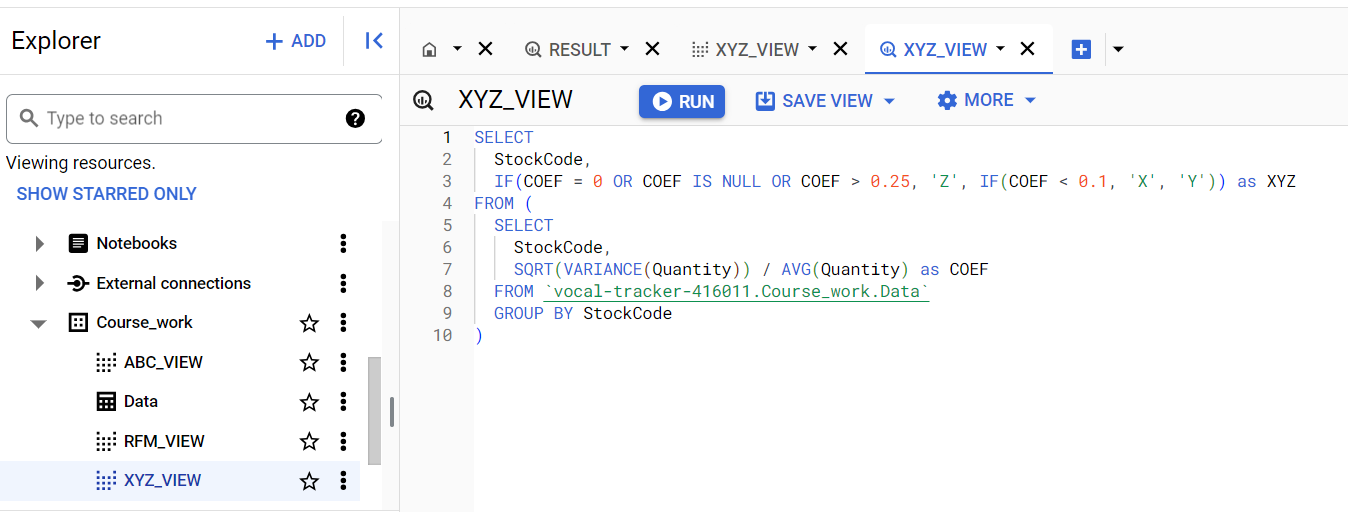


Рисунок 10 - XYZ анализ в виде представления VIEW в BigQuery

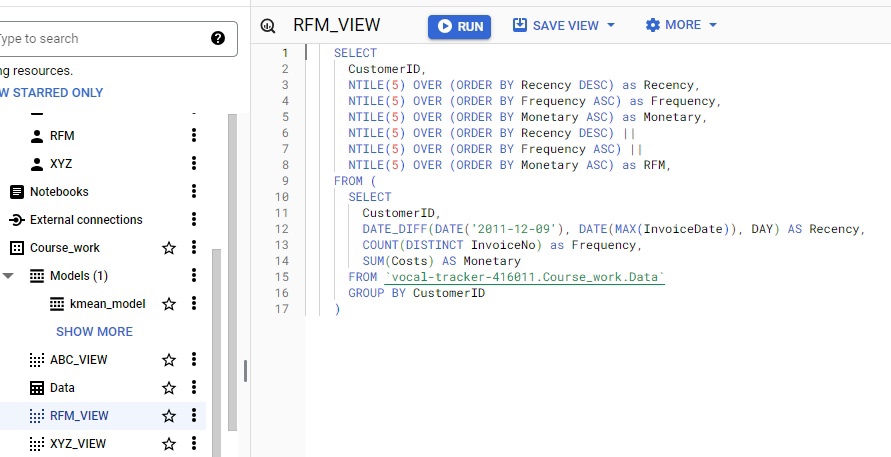


Рисунок 11 - RFM анализ в виде представления VIEW в BigQuery

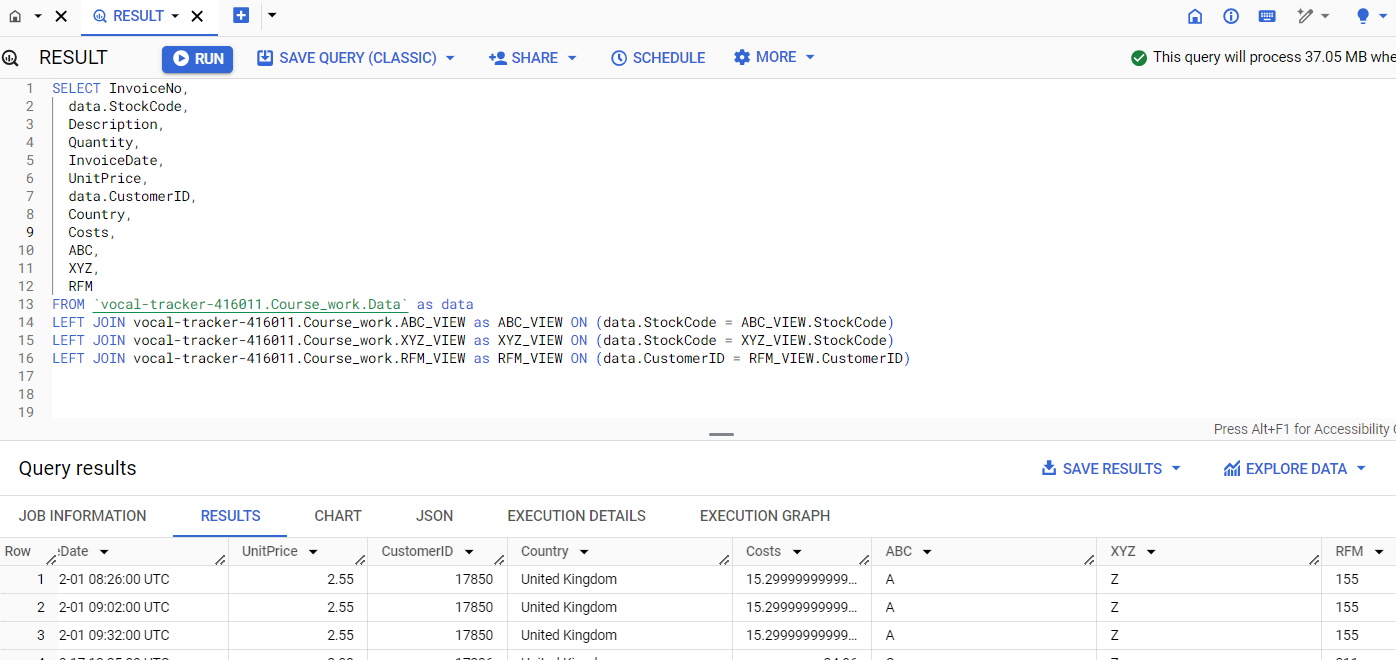


Рисунок 12 - Итоговый результат в BigQuery

Как видно, данные, полученные SQL-запросами на платформе Google BigQuery совпадают с аналогичным результатом в Google Colab.

# **2.3. Загрузка и анализ данных в Loginom**

Вначале было настроение подключение из BigQuery и проведен аналогичный анализ данных. На рисунке 13 представлена вся схема исследовательского анализа.

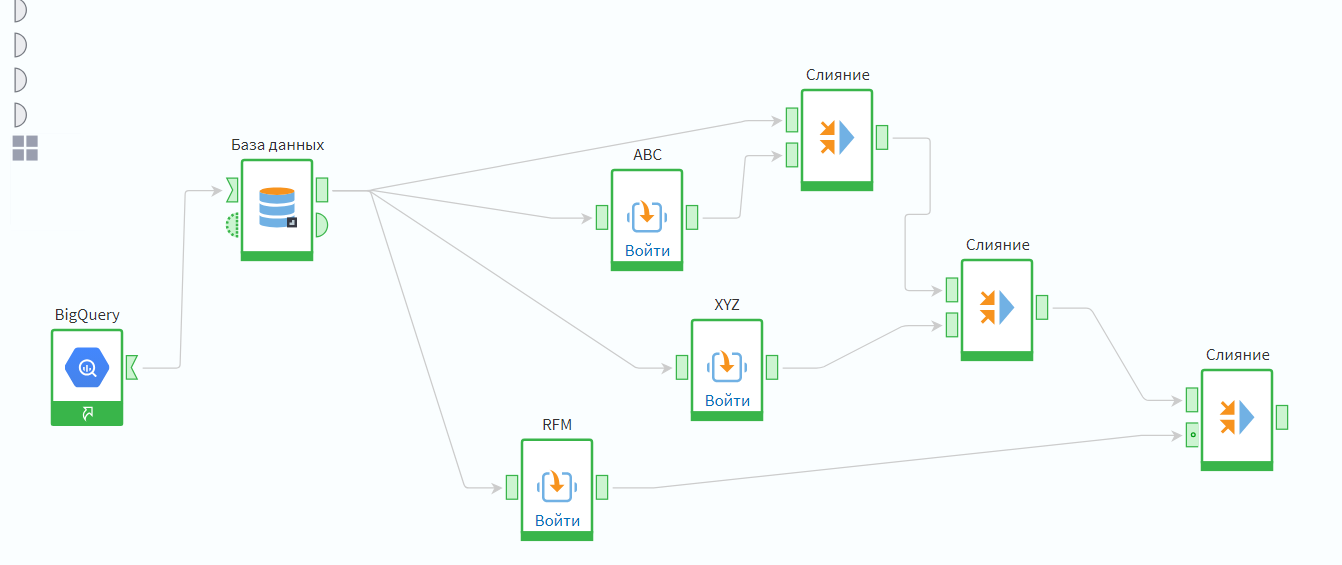


Рисунок 13 - Анализ данных в Loginom

Для начала был создан сервисный аккаунт и скачан приватный ключ в формате json (см. Рисунок 14). Далее создано подключение в Loginom и загружен датасет с помощью модуля “База данных”. После этого по были проведены ABC, XYZ и RFM анализы по такой же логике, как и в предыдущих разделах. После каждого из этапов результаты анализа были соединены с изначальной витриной данных.

Для ABC анализа в начале была произведена группировка по продуктам (см. Рисунок 15). Далее данные были отсортированы по суммарному доходу с товара в порядке убывания (см. Рисунок 16). После чего высчитан процент от общего дохода (см. Рисунок 17) и накопительная сумма (см. Рисунок 18). После чего проведена сегментация по уже известному условию на три группы A, B и C (см. Рисунок 19). Данные расчеты позволяют делать модуль «Калькулятор». Все используемые модули и их связи показаы на рисунке 20.

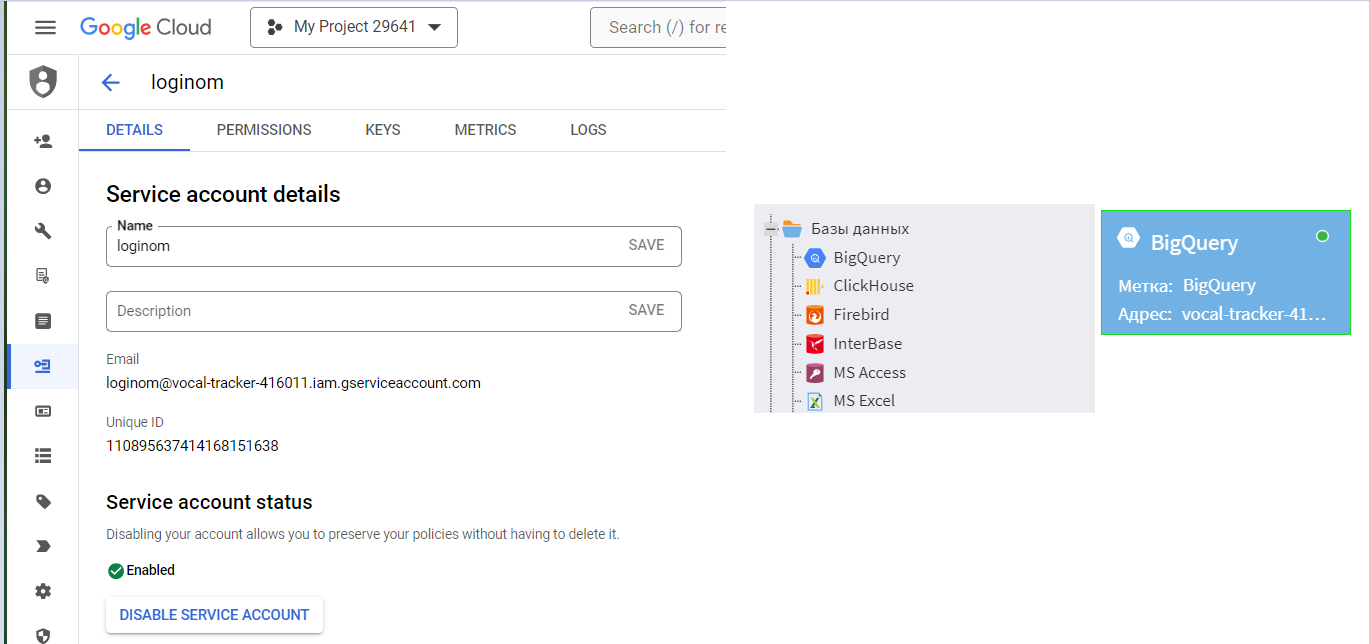


Рисунок 14 - Подключение к облачному хранилищу BigQuery

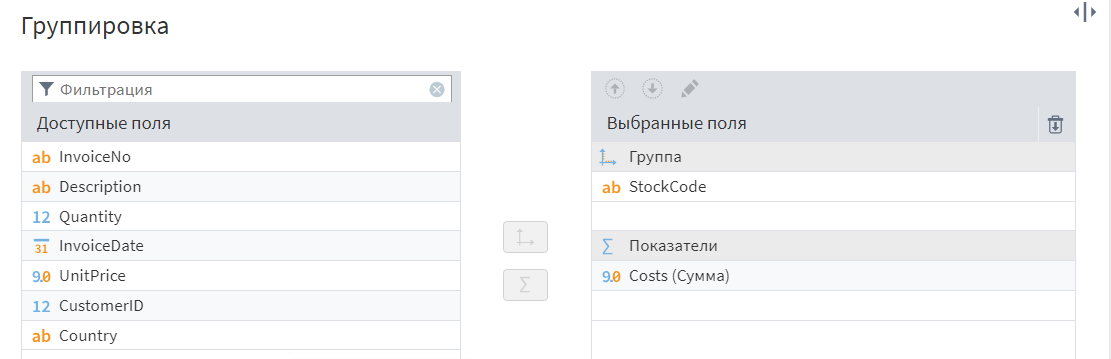


Рисунок 15 - Группировка по StockCode в Loginom

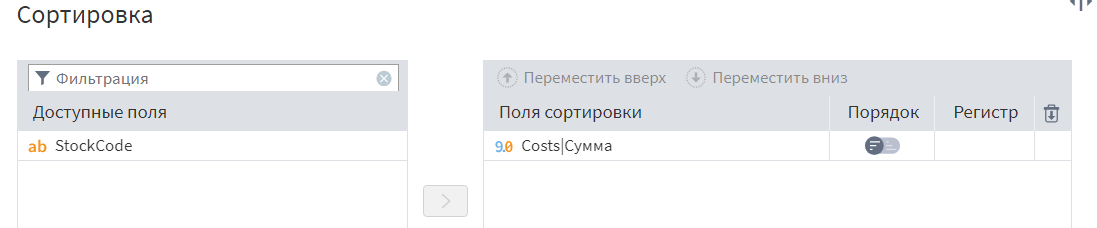


Рисунок 16 - Сортировка в порядке убывания в Loginom

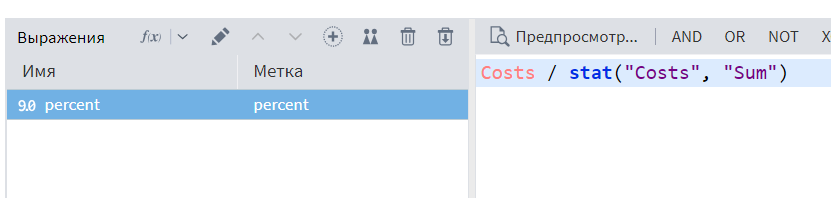


Рисунок 17 - Процент от общего дохода в Loginom

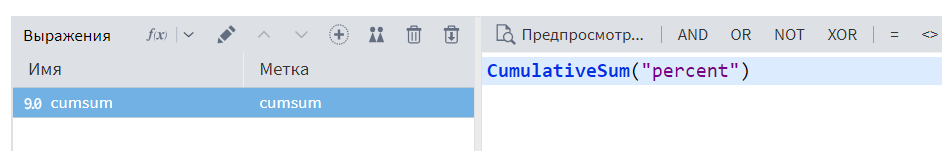


Рисунок 18 - Накопительная сумма в Loginom

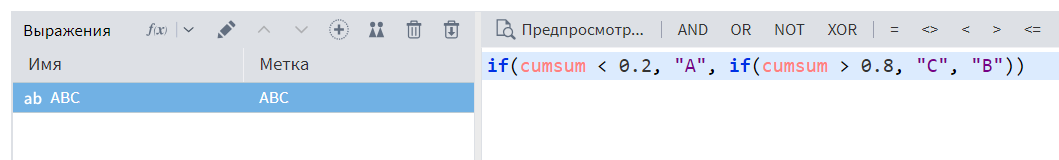


Рисунок 19 - Сегментация товара в ABC анализе в Loginom

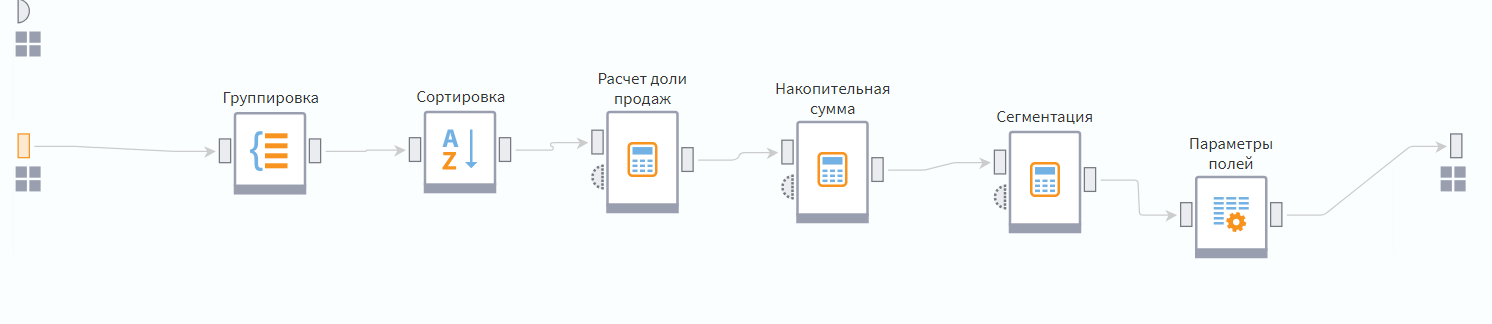


Рисунок 20 - Общая схема ABC анализа в Loginom

Для проведения XYZ анализа была произведена группировка по товару (см. Рисунок 21). Затем рассчитан коэффициент вариации (см. Рисунок 22) и проведена сегментация (см. Рисунок 23). Общая схема XYZ анализа в Loginom приведена на рисунке 24.

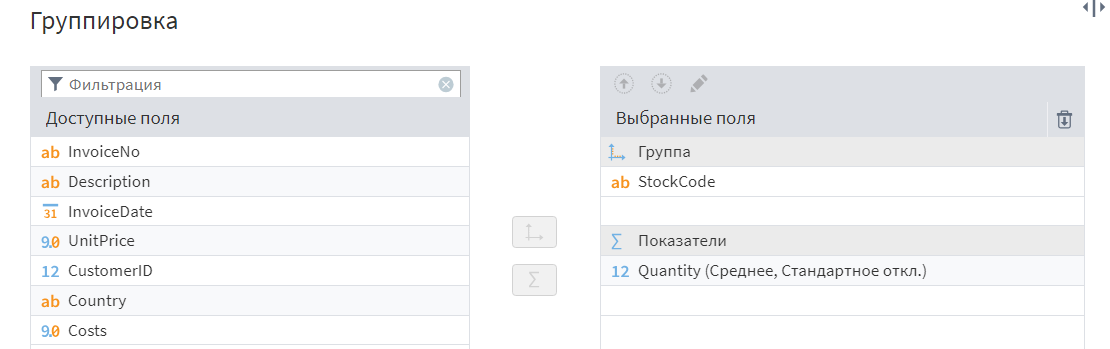


Рисунок 21 - Группировка по товару для XYZ

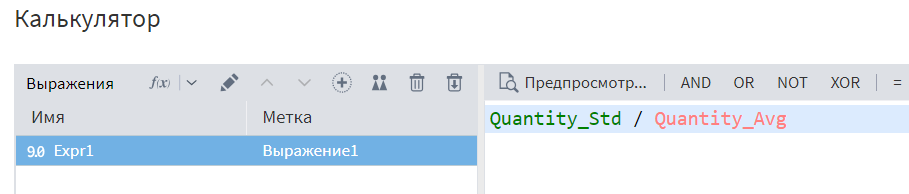


Рисунок 22 - Расчет коэффициента вариации в Loginom

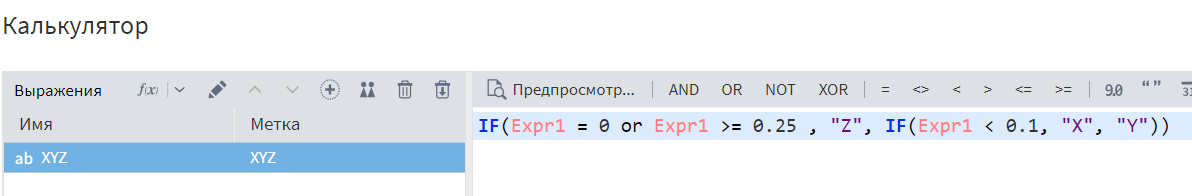


Рисунок 23 - Сегментация товаров для XYZ анализа в Loginom

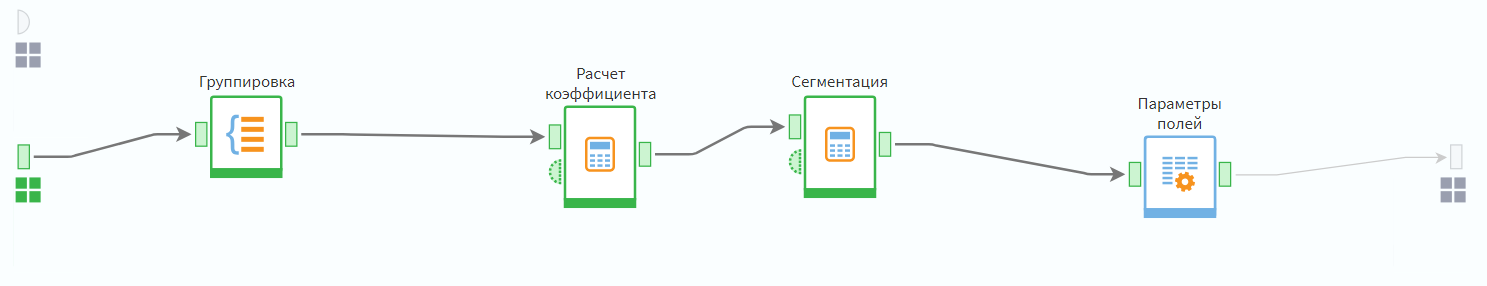


Рисунок 24 - Общая схема XYZ анализа в Loginom

Для проведения RFM анализа таблица была сгруппирована по коду клиента (см. Рисунок 25). С помощью функции «DaysBetween» рассчитан показать Recency – Давность (см. Рисунок 26). Показатели Monetary и Frequency рассчитаны сразу на этапе группировки, как общая сумма дохода по клиенту и количество уникальных сделок соответственно. После чего три показателя разбивались на 5 групп по квантилям с помощью метода квантования «Плитка» (см. Рисунок 27). Далее три метки групп соединены в один столбец с помощью конкатенации строк «Concat» (см. Рисунок 28). Общая схема RFM анализа приведена на рисунке 29.

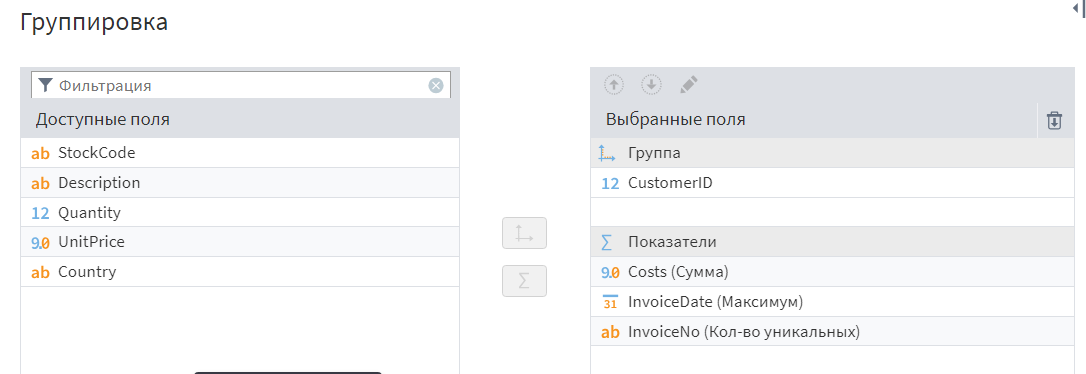


Рисунок 25 - Группировка по клиенту для RFM анализа в Loginom

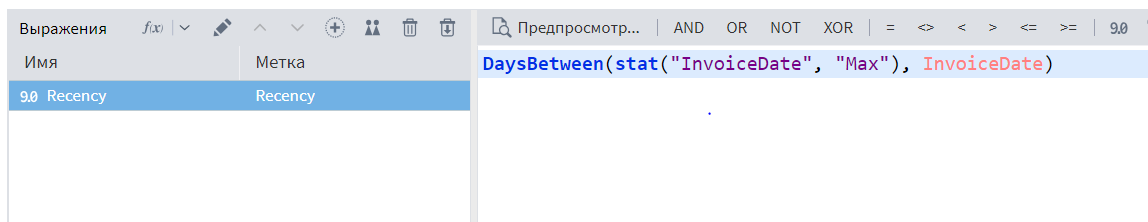


Рисунок 26 - Расчет Recency для RFM анализа в Loginom

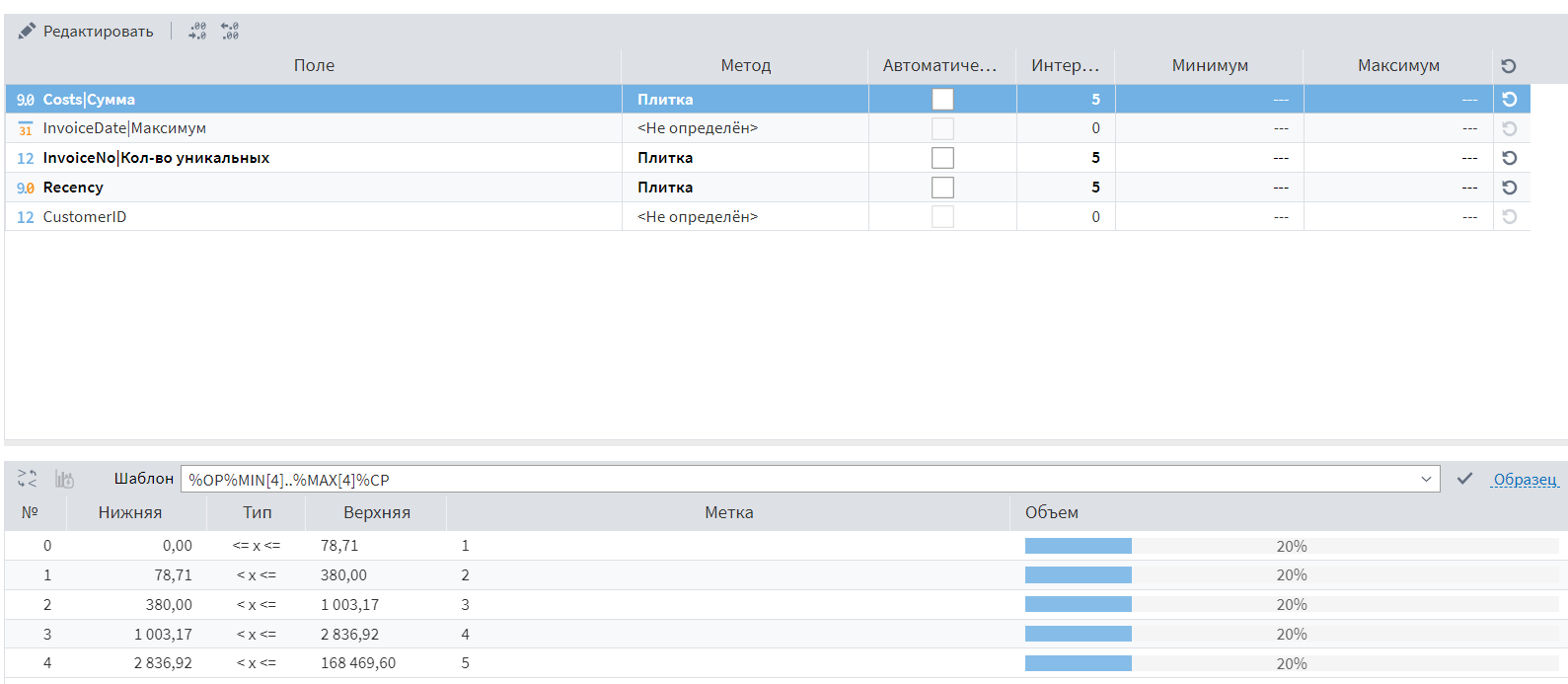


Рисунок 27 - Квантование для RFM анализа в Loginom

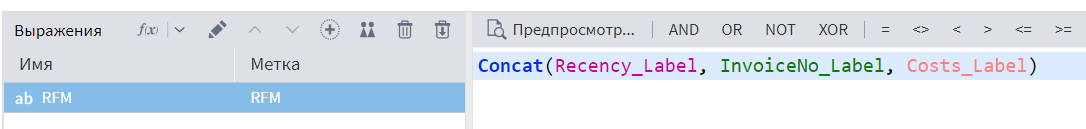


Рисунок 28 - Конкатенация строк в Loginom

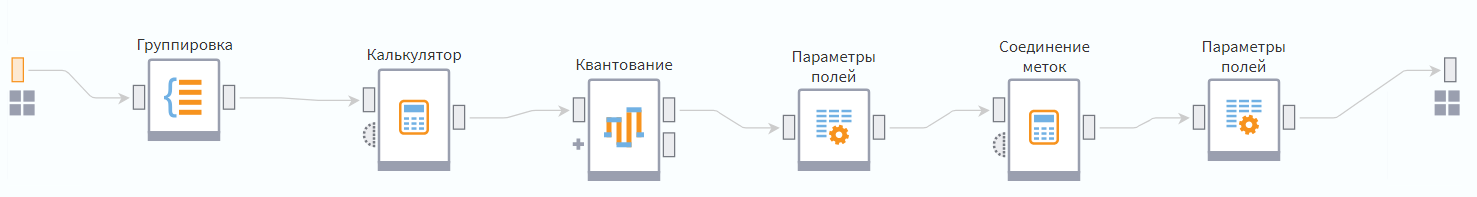


Рисунок 29 - Общая схема RFM анализа в Loginom.

Результаты анализов представлены на рисунке 30. Как видно, результат идентичен с другими платформами.

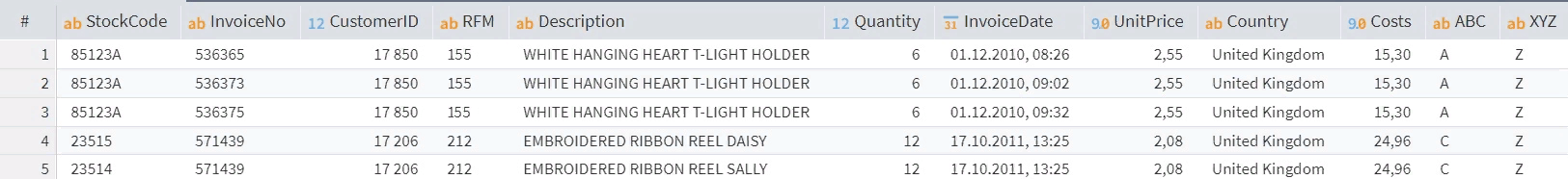


Рисунок 30 - Результаты ABC, XYZ, RFM анализов в Loginom

# **2.4. Кластеризация клиентов с применением машинного обучения в BigQuery.**

Помимо приведенных выше возможностей, облачная платформа BigQuery - одна из немногих, которая позволяет использовать основные методы машинного обучения на языке SQL. Для кластеризации клиентов на 4 группы был применен алгоритм классического машинного обучения K-means. Для этого при помощи уже готового представления из пункта 2.2 была создана модель машинного обучения при заданном числе кластеров с помощью метода CREATE MODEL (см. Рисунок 31, 32). Затем на основе созданной модели при помощи метода ML\_PREDICT было получено разбиение клиентов на 4 кластера (см. Рисунок 33). Здесь номер кластера отражает поле CENTROID\_ID.

По координатам центров кластеров можно понять, что 4 кластер объединяет относительно новых клиентов. Последний заказ у них был недавно, частотность и прибыль пока невысокие. Кластер 2 объединяет самых лояльных клиентов с высокой прибылью и частотой. Кластер 3 включает в себя постоянных клиентов. А вот кластер 1 объединяет клиентов в зоне риска. Последний заказ был достаточно давно, низкая частотность и низкая выручка.

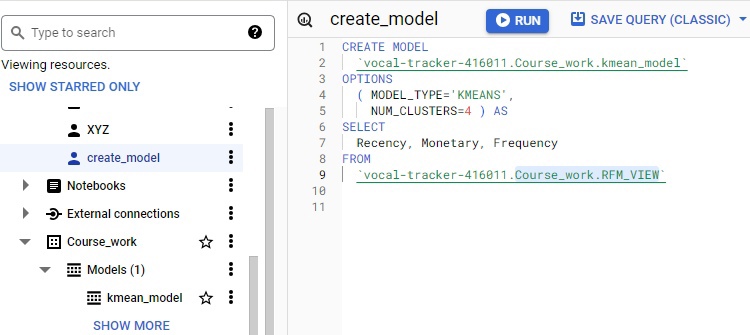


Рисунок 31 - Создание модели кластеризации в BigQuery

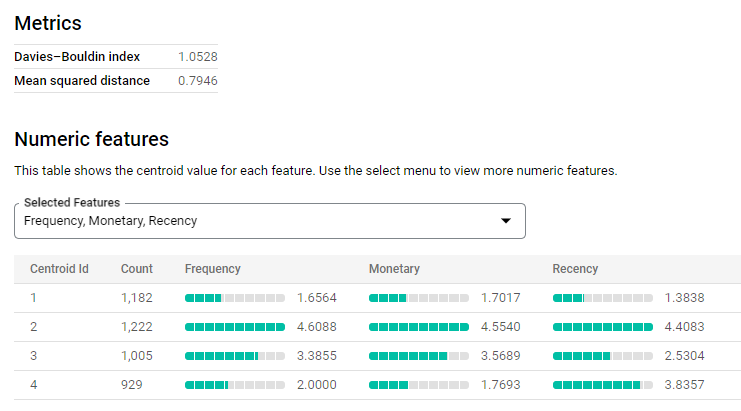


Рисунок 32 - Параметры полученной модели

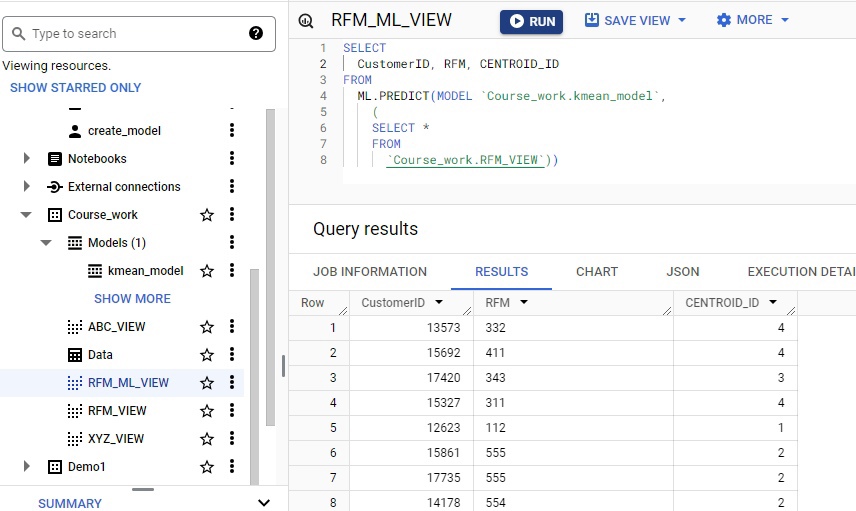


Рисунок 33 - Результат работы модели кластеризации в BigQuery

# **ЗАКЛЮЧЕНИЕ**

В ходе проделанной работы были изучены теоретические основы применения облачных технологий в продуктовой аналитике. Также проведены ABC и XYZ анализы по товарам, что дает возможность разбить их на группы и выстроить грамотную стратегию бизнеса, понять на какие товары стоит сделать акцент, а какие товары стоит исключить. Проведен RFM анализ по клиентам с целью из разбиения на группы, что дает возможность проанализировать наш продукт как со стороны клиентов, так и со стороны товаров. При проведении исследований использованы облачные технологии Google BigQuery в качестве как хранилища данных, так и средства построения аналитик, Google Colab для аналитики и предварительной обработки перед загрузкой в хранилище. Также продемонстрировано подключение к облачному хранилищу из локальной программы Loginom. Все анализы были проведены на трех платформах и получены одинаковые результаты. Дополнительно была построена модель машинного обучения для кластеризации клиентов на основе показателей RFM анализа в платформе Google BigQuery.

# **СПИСОК ЛИТЕРАТУРЫ**

1. Ojala A. Discovering and creating business opportunities for cloud services //Journal of Systems and Software. – 2016. – Т. 113. – С. 408-417.
2. Górski T., WOźniak A. P. Optimization of business process execution in services architecture: A systematic literature review //IEEE Access. – 2021. – Т. 9. – С. 111833-111852.
3. Паус А. С., Целовальникова О. А. Тенденции развития облачных технологий на российском рынке //Новые информационные технологии в автоматизированных системах. – 2014. – №. 17. – С. 486-492.
4. Ng W. L. A simple classifier for multiple criteria ABC analysis //European Journal of Operational Research. – 2007. – Т. 177. – №. 1. – С. 344-353.
5. Pandya¹ B., Thakkar H. A review on inventory management control techniques: ABC-XYZ analysis //REST Journal on Emerging trends in Modelling and Manufacturing. – 2016. – Т. 2. – №. 3. – С. 2016.
6. Christy A. J. et al. RFM ranking–An effective approach to customer segmentation //Journal of King Saud University-Computer and Information Sciences. – 2021. – Т. 33. – №. 10. – С. 1251-1257.
7. Яковлев В. Б. Анализ данных в аналитической платформе Loginom. – 2020.
8. Ткаченко А. Л., Мельников А. А., Кузнецова В. И. Прикладные решения на базе Loginom //Дневник науки. – 2021. – №. 5.
9. Ali M. H., Hosain M. S., Hossain M. A. Big Data analysis using BigQuery on cloud computing platform //Australian JofEng Inno Tech. – 2021. – Т. 3. – №. 1. – С. 1-9.
10. Tigani J., Naidu S. Google bigquery analytics. – John Wiley & Sons, 2014.