МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯРОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Уральский государственный экономический университет»
(УрГЭУ)

Направление подготовки	38.03.05 Бизнес - информатика
Направленность (профиль)	Электронный бизнес

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА (БАКАЛАВРИАТ)

Тема	**Интеллектуальный анализ продаж на производственном предприятии**

Обучающийся	**Фролов Максим Андреевич**
Группа	БИ-16
Руководитель	Галактионов Александр Дмитриевич
	доцент, к.х.н.

Кафедра	Бизнес - информатика
Институт	Менеджмента и информационных технологий

Нормоконтролер	Бегичева Светлана Викторовна
	старший преподаватель

Дата защиты	26.06.2020
Оценка	

Екатеринбург

2020 г.

СОДЕРЖАНИЕ

Список сокращений.....3

Введение.....4

1 Теоретические аспекты интеллектуального анализа данных.....7

1.1 Основные понятия Data Mining.....7

1.2 Анализ методов Data Mining.....9

1.3 Обзор инструментальных средств интеллектуального анализа данных…12

1.3.1 QlikView.....15

1.3.2 Power BI.....16

1.3.3 Tableau.....18

2 Применение неиерархического метода кластеризации и методов визуализации для интеллектуального анализа данных средствами Tableau.....21

2.1 Сущность алгоритмов кластеризации и методов визуализации в анализе данных.....21

2.2 Процесс кластеризации в среде Tableau.....31

2.3 Процесс визуализации в среде Tableau.....33

3 Интеллектуальный анализ продаж в среде Tableau.....37

3.1 Подключение источника данных.....37

3.2 Кластерный анализ продаж.....41

3.3 Анализ динамики продаж.....55

3.4 Прогнозирование.....62

3.5 Комплексный анализ продаж.....65

Заключение.....69

Список использованных источников.....71

СПИСОК СОКРАЩЕНИЙ

SAS	Statistical Analysis System
IBM	International Business Machines
SQL	Structured Query Language
SAP	Systems, Applications and Products
ETL	Extract, Transform, Load
BI	Business intelligence
ПО	Программное обеспечение
ОС	Операционная система

ВВЕДЕНИЕ

За последние пару десятилетий мы стали свидетелями повсеместного распространения новейших информационных технологий и их внедрения в различные области социальной и экономической сфер. Всё это вкупе с наличием свободного доступа к сети Интернет привело к огромному росту объёма информации, производимой человеком.

Только за пару дней человечество сегодня производит столько информации, сколько было создано за всё время существования Земли вплоть до нынешнего тысячелетия. По утверждениям исследовательских компаний, сейчас объём данных на планете удваивается каждые пару лет.

Из-за огромного количества существующей информации, лишь малая её часть будет увидена человеческим глазом, и, как следствие, значительные объёмы практичной и полезной информации просто пройдут мимо. Это приводит нас к пониманию важности проблем, сопряжённых с анализом накопленных данных для извлечения новых знаний.

Наша единственная возможность понять и найти что-то полезное в таком многообразии данных – это широкое применение методов интеллектуального анализа данных. Интеллектуальный анализ данных широко используется во многих областях с большим объёмом данных. В науке – биологии, медицине, астрономии и других сферах деятельности. В бизнесе – в первую очередь в торговле, телекоммуникации, промышленном производстве и других.

Актуальность использования интеллектуального анализа данных в бизнесе связана с жесткой конкуренцией, возникшей вследствие перехода от «рынка производителя» к «рынку потребителя». В этих условиях особенно важно качество и обоснование принимаемых решений, что требует строгого количественного анализа имеющихся данных. При работе с большими объемами накапливаемой информации необходимо постоянно оперативно отслеживать динамику рынка, а это практически невозможно без автоматизации аналитической деятельности.

В настоящее время проблема контроля финансов на промышленном предприятии является очень серьезной. Можно сказать, что современные системы управления финансами необходимо постоянно обновлять после стадии разработки, начиная с простейшего расчета резервов и производственных мощностей вручную и заканчивая новейшими компьютерными методами автоматизации. Поскольку современная экономика выступает за совершенно иную идею управления промышленными предприятиями, необходимо пересмотреть научные и практические стратегические вопросы повышения эффективности управления финансовой стабильностью.

Объектом исследования: методы и инструментальные средства интеллектуального анализа данных.

Предметом исследования является методы кластеризации и визуализации в анализе продаж.

Целью выпускной квалификационной работы является реализация применения методов кластеризации и визуализации для интеллектуального анализа продаж на производственном предприятии.

Для реализации указанной цели следует решить ряд задач:

- рассмотреть технологию интеллектуального анализа и его методы;
- изучить и выбрать инструментальные средства;
- рассмотреть использование методов, в выбранном инструментальном средстве;
- провести анализ продаж на производственном предприятии, используя методы интеллектуального анализа.

Выпускная квалификационная работа состоит из введения, трёх глав, заключения и списка использованных источников. Включает в себя 68 рисунков.

Информационную базу выпускной квалификационной работы составляет выборка из переработанных данных продаж предприятия «Coca-Cola HBC Eurasia».

Выпускная квалификационная работа состоит из введения, трёх глав, заключения и списка использованных источников.

Во введении обоснована актуальность темы, определены: объект, предмет, цель и задачи исследования.

В первой главе рассматриваются теоретические основы интеллектуального анализа, методы и алгоритмы, и инструментальные средства.

Во второй главе рассмотрена сущность алгоритмов кластеризации и методов визуализации, а также их использование в среде Tableau.

В третьей главе проведён интеллектуальный анализ продаж на производственном предприятии в среде Tableau.

1 ТЕОРЕТИЧЕСКИЕ АСПЕКТЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

1.1 ОСНОВНЫЕ ПОНЯТИЯ DATA MINING

Data Mining имеет множество понятий, часть из которых мы рассмотрим ниже.

Data Mining — мультидисциплинарная область, возникшая и развивающаяся на базе таких наук как прикладная статистика, распознавание образов, искусственный интеллект, теория баз данных, алгоритмизация, визуализация и других.

Data Mining — это процесс поддержки принятия решений, основанный на поиске в данных скрытых закономерностей.

Data Mining — это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности (Григорий Пиатецкий-Шапиро) [7].

Суть и цель технологии Data Mining можно охарактеризовать так: это технология, которая предназначена для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей [1].

Неочевидных – от того, что стандартные пути в выявлении закономерности бесполезны.

Объективных – это означает, что обнаруженные закономерности будут соответствовать действительности и при этом не будут субъективными как это, было бы, будь они экспертными [3].

Практически полезных – что означает применимость последующего результата в реальной жизни на практике.

Знания – это определенный набор данных о том или ином предмете или области, знания могут быть ложными и истинными. Ложные знания подразумевают под собой недостоверные сведенья, а истинные напротив

является достоверными. Знания должны описывать новые связи между свойствами, предсказывать значения одних признаков на основе других.

Существуют и другие определения понятия Data Mining.

Data Mining — это процесс выделения, исследования и моделирования больших объемов данных для обнаружения неизвестных до этих структур с целью достижения преимуществ в бизнесе (определение SAS Institute) [Error: Reference source not found

Data Mining – это поиск скрытых закономерностей, выявление причинно-следственных связей, анализ рисков.

Data Mining – объединение различных механизмов, ориентированных на решение 5 классов задач: регрессия, классификация, кластеризация, ассоциация, последовательность.

- регрессия – установление зависимости непрерывных выходных переменных от входных значений.
- классификация – это отнесение объектов к одному из заранее известных классов.
- кластеризация – объекты внутри кластера должны быть «похожими» друг на друга и отличаться от объектов, вошедших в другие кластеры.
- ассоциация – нахождение зависимости, что из события X следует событие Y.
- последовательность – установление зависимостей между связанными во времени событиями [Error: Reference source not foundError: Reference source not found].

Data Mining – один из немногих способов, с помощью которых можно прогнозировать и моделировать различные процессы, что позволяет нам проводить эксперименты с различными бизнес-процессами в компании без рисков потерять средства. Обычно в компаниях накапливаются огромные массивы данных, на которые при ручном расчете хотя бы одной ситуации уйдут годы или даже десятилетия [2].

В данный момент Data Mining используются практически во всех сферах деятельности человека из-за его огромных преимуществ.

Рисунок 1 иллюстрирует некоторые популярные бизнес-задачи, которые решаются алгоритмами Data Mining.

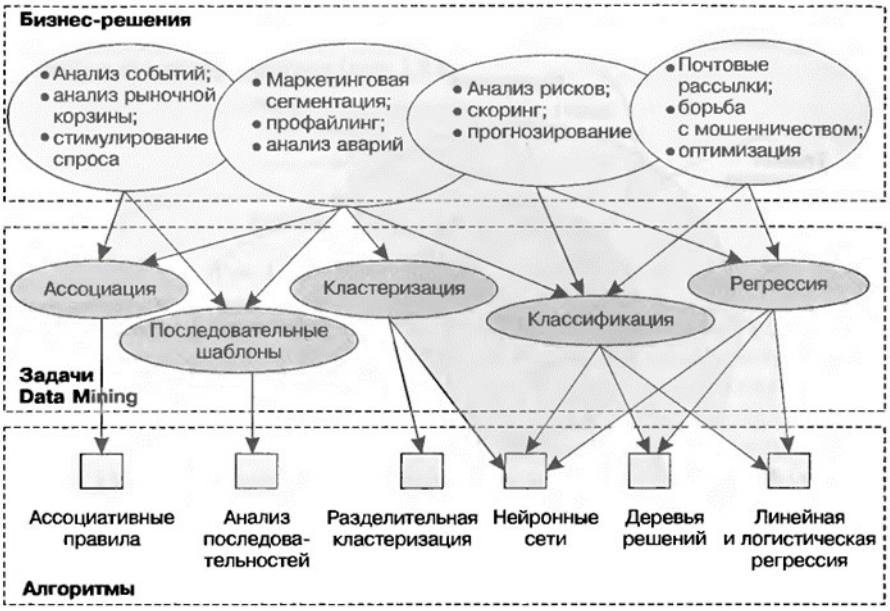

Рисунок 1 – бизнес-задачи, решаемые алгоритмами Data Mining [4]

1.2 АНАЛИЗ МЕТОДОВ DATA MINING

Новый взгляд на бизнес-процессы – следствие использования методов интеллектуального анализа, который позволяет увидеть не обнаруженные вручную скрытые закономерности и свойства у исследуемых процессов и объектов.

Метод представляет собой норму или правило, определенный путь, способ, прием решений задачи теоретического, практического, познавательного, управленческого характера [1].

Алгоритм – описание последовательных действий, направленных на получение искомого результата.

Методы Data Mining можно классифицировать на статистические и кибернетические [5].

Статистические методы Data Mining можно классифицировать на 4 группы:

- дескриптивный анализ и описание исходных данных;
- анализ связей (корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ);
- многомерный статистический анализ (компонентный анализ, дискриминантный анализ, многомерный регрессионный анализ, канонические корреляции и др.);
- анализ временных рядов (динамические модели и прогнозирование).

Кибернетические методы Data Mining

Второе направление Data Mining – это множество подходов, объединенных идеей искусственных компьютерных сетей.

К этой группе относятся такие методы:

- искусственные нейронные сети (распознавание, кластеризация, прогноз);
- эволюционное программирование (в том числе алгоритмы метода группового учета аргументов);
- генетические алгоритмы (оптимизация);
- ассоциативная память (поиск аналогов, прототипов);
- нечеткая логика;
- деревья решений;
- системы обработки экспертных знаний [5].

Нейронные сети являются самообучающимся алгоритмом, учитывающим взаимное влияние вклада каждого фактора в финальный результат, который определяется автоматически.

Нейронные сети основаны на имитации человеческого мозга. Они способны не только выполнять однажды запрограммированную последовательность действий над заранее определенными данными, но и сами анализировать вновь поступающую информацию. Эффективная постройка нелинейных зависимостей, которые более точно описывают наборы данных по сравнению с линейными методами постройки, являются основным достоинством нейронных систем [13].

Основными направлениями использования нейронных сетей являются: распознавания образов, прогнозирование, адаптивное управление, создание экспертных систем, организация ассоциативной памяти, обработка аналоговых и цифровых сигналов.

Деревья решений – это древовидная иерархическая модель, где в каждом узле производится проверка определенного атрибута (признака) с помощью правила. Отображена на рисунке 2.

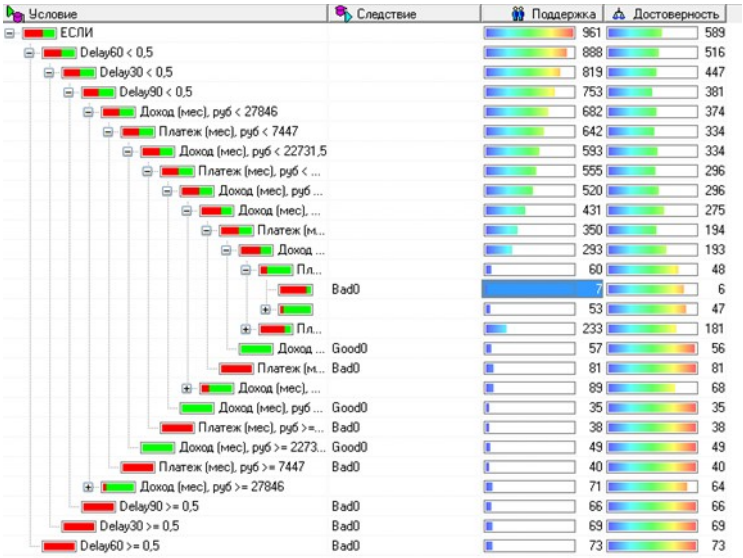

Рисунок 2 – дерево решений

По результатам проверки формируются два или более дочерних узла, в которые попадают объекты, для которых значения данного атрибута удовлетворяют (или не удовлетворяют) правилу в родительском узле.

Под правилом понимается логическая конструкция, представленная в виде «если ..., то ...» [10].

Достоинствами алгоритма является простота интерпретации полученных результатов и автоматический выбор алгоритмом наиболее значимых факторов.

Недостатком является низкая точность данных.

С помощью метода дерево решений решаются задачи классификации и прогнозирования.

Если зависимая, т.е. целевая переменная принимает дискретные значения, при помощи метода дерева решений решается задача классификации.

Если же зависимая переменная принимает непрерывные значения, то дерево решений устанавливает зависимость этой переменной от независимых переменных, т.е. решает задачу численного прогнозирования.

Логистическая регрессия – разновидность множественной регрессии, предназначенная для классификации записей на основании значений входных полей. При этом выходная переменная является категориальной или бинарной.

В бинарной классификации каждый объект или наблюдение должны быть отнесены к одному из двух классов (например, А и Б). Тогда с каждым исходом связано событие: объект принадлежит к классу А и объект принадлежит к классу Б. Результатом будет оценка вероятности соответствующего исхода. Если в процессе анализа будет установлено, что вероятность принадлежности объекта с заданным набором значений признаков (входных переменных) к классу А больше, чем вероятность его принадлежности к классу Б, то он будет классифицирован, как объект класса А.

Логистическую регрессию и ROC-анализ почти всегда можно увидеть в наборе Data Mining алгоритмов, поскольку является эффективным инструментом оценки качества моделей [11].

1.3 ОБЗОР ИНСТРУМЕНТАЛЬНЫХ СРЕДСТВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

На рынке программного обеспечения Data Mining существует огромное разнообразие продуктов, относящихся к этой категории. И не растеряться в нем достаточно сложно. Для выбора продукта следует тщательно изучить задачи, поставленные перед аналитиком, и обозначить те результаты, которые необходимо получить [14].

На рисунке 3 представлены основные сегменты рынка программного обеспечения в области интеллектуального анализа данных.

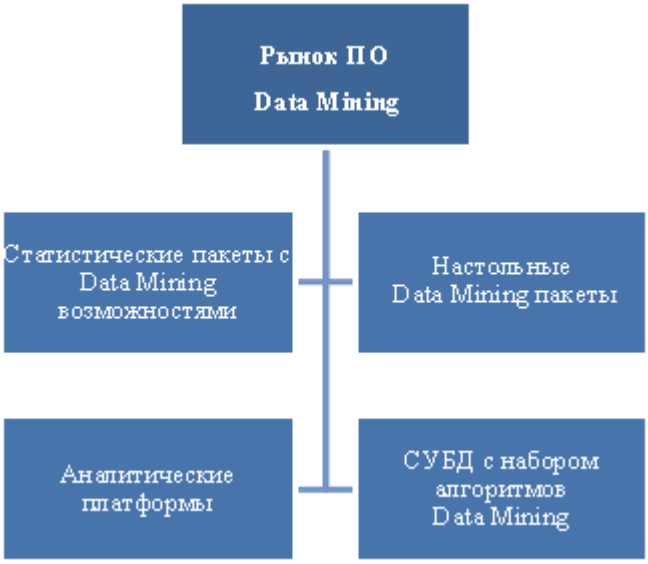

Рисунок 3 – рынок программного обеспечения в области интеллектуального анализа данных

Статистические пакеты с возможностями Data Mining и настольные Data Mining пакеты ориентированы в основном на профессиональных пользователей. Их отличительные особенности [Error: Reference source not found]:

- слабая интеграция с промышленными источниками данных;
- бедные средства очистки, предобработки и трансформации данных;
- отсутствие гибких возможностей консолидации информации, например, в специализированном хранилище данных;
- конвейерная, поточная обработка новых данных затруднительна или реализуется встроенными языками программирования и требует высокой квалификации;
- из-за использования пакетов на локальных рабочих станциях обработка больших объемов данных затруднена.

Статистические пакеты широко распространены. В зависимости от постановки задачи настольные пакеты интеллектуального анализа могут быть направлены как на решение одной задачи, например кластеризация, так и всех классов задач. Обширные возможности алгоритмов, предоставляемые данными пакетами, достаточны для решения исследовательских задач. Существует немало свободно распространяемых настольных пакетов Data Mining с открытыми исходными кодами.

В качестве примеров распространенных статистических пакетов можно назвать: SAS от компании SAS Institute, IBM SPSS STATISTICS компании IBM и STATISTICA компании Statsoft. Сегмент настольных пакетов Data Mining представлен такими программами как пакет R, WEKA, KNIME, RapidMiner.

Для прикладных решений промышленного уровня используют аналитические платформы, совместно с СУБД, в которой есть возможности применения методов data mining. Такое решение эффективнее, нежели использование статистических пакетов, поскольку последние в данном случае неэффективны. На сегодняшний день все крупные производители СУБД включают в состав своих продуктов поддержку хранилищ данных вместе с средствами анализа данных. Это позволяет максимально эффективно использовать алгоритмы анализа данных и держать производительность на высоком уровне. Однако при таком решении появляется жёсткая привязка всех технологий анализа к одной СУБД, а также добавляет сложность в создании прикладных решений из-за ориентированности СУБД на программистов и администраторов баз данных.

Представителями данного класса систем являются Microsoft SQL Server Analysis Services от Microsoft Corporation, Oracle Business Intelligence компании Oracle Corporation, IBM DB2 Intelligent Miner компании IBM.

Аналитические платформы напротив, сразу направлены на анализ данных и предназначены для создания готовых решений. В аналитической платформе, как правило, всегда присутствуют гибкие и развитые средства консолидации, включающие богатые механизмы интеграции с промышленными источниками данных, инструменты очистки и преобразования структурированных данных и их последующее хранение в едином источнике в многомерном виде — в хранилище данных. Модели, описывающие выявленные закономерности, правила и прогнозы, также хранятся в специальном источнике данных — репозитарии моделей.

Сегодня аналитические платформы — это самый быстрорастущий и привлекательный сегмент рынка программного обеспечения интеллектуального

анализа данных. Лидерами являются компании: SAS, SAP, IBM, Oracle, Microstrategy, Microsoft, Information Builders.

Средства графического отображения данных, специализирующиеся на функции визуализации, занимают значительную долю рынка. Примером могут служить следующие компоненты и продукты: Qlik Sense, Power BI, Tableau.

В данных системах преобладает удобство и простота использования интерфейса, с помощью которого возможно быстро составлять нужные графики, ассоциируя с анализируемыми показателями различные параметры диаграммы. К таким параметрам относятся цвет, форма, ориентация относительно собственной оси, размеры и другие свойства графических элементов изображения. Дополнительно средства визуализации данных снабжены удобными средствами для масштабирования и вращения изображений, в том числе на различные мобильные программные платформы.

Инструментальное средство будет выбираться из группы аналитической платформы, специализирующееся на графическом отображении данных.

1.3.1 QlikView

QlikView — платформа, которая концентрируется на пользователе, как на получателе данных. Она позволяет ему исследовать свои данные по тем же принципам, что и разработчик, который занимается их обработкой. При этом, чтобы поддерживать гибкость в своем подходе к изучению и визуализации данных, система сохраняет связи между данными. Это позволяет пользователю вместе с куском данных, который он хотел найти, достать и другую релевантную информацию, которая также будет полезна [21].

QlikView есть возможность тонкой настройки любого объекта, благодаря чему можно настроить графики и дашборды так, как нужно пользователю.

Собственная разработка – ассоциативная модель данных, позволяет сохранять таблицы в оперативной память, благодаря чему данные обрабатываются с высокой скоростью, а любое значение из одной таблицы со

всеми значениями из других таблиц. Внутрь QlikView встроен ETL, позволяющий обрабатывать и очищать данные.

С помощью большого количества функций, при правильной настройке, возможно создать продвинутые дашборды на основе данных из разных источников. Например, есть такая функция, ускоряющая процесс создания дашбордов и отчётов, за счёт исключения человеческого вмешательства, а именно самостоятельного автоматического распознавания связей между данными.

Также присутствует техника сжатия данных до десяти процентов от изначального размера, что позволяет увеличить скорость обработки и анализа данных и сократить количество места, требуемого оперативной памятью [20].

Для уверенной работы в QlikView, необходимо изучение собственного языка данной программы, и продвинутое умение писать SQL запросы, поскольку отчёты и графики создаются не просто.

Стоимость подписки у данного продукта одна из самых высоких и запутанных, в отличие от конкурентов.

Плюсы и минусы QlikView отображены на рисунке 4:

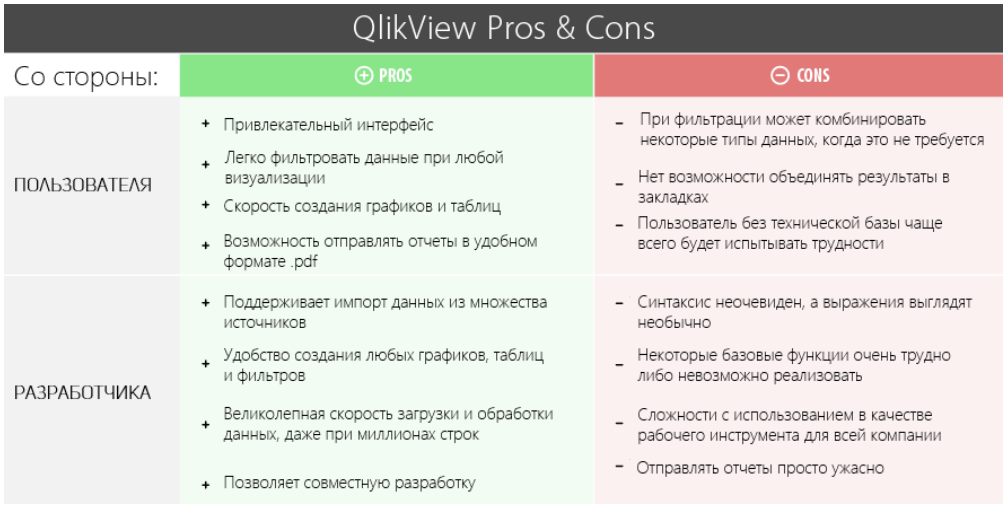

Рисунок 4 – плюсы и минусы QlikView

1.3.2 Power BI

Power BI — это онлайн-сервис, разработанный Microsoft для бизнес-аналитики с возможностью подключения различных источников данных и

сторонних приложений. Платформа обладает веб-интерфейсом, позволяющим создавать кастомизированные визуализации, а с помощью настольного приложения можно проводить стандартизацию и очистку данных. Также существует мобильная версия Power BI, доступная на различных ОС, чтобы принимать решения на ходу [19].

Из особенностей Power BI отметим стилизованный под Excel интерфейс со своими доработками, который изначально планировался расширить функциональность MS Excel. Тесная связь с главными продуктами компании обусловлена принадлежностью к Microsoft [8].

Бесплатная базовая версия присутствует, срок использования которой составляет 60 дней. Присутствует поддержка большого количества способов импорта данных, простое API для интеграции в другие приложения, возможность наблюдать в реальном времени за изменением данных на интерактивных дашбордах м возможностью делиться ими разными способами. Из приятных особенностей – присутствие кроссплатформенности.

Удобство использования обусловлено простотой создания визуализаций, где нужно лишь переместить необходимый элемент на пустое место в отчёте или дашборде. Аналогично и с данными. Такой метод работы с данными называется drag-and-drop, что в переводе на русский язык означает перетащи и брось.

Присутствует бесплатная версия с ограничениями на лимит памяти в 1 гигабайт и скоростью обработки потоковых данных до 10 тыс. строк в час. Обновление и совместная работа над отчётами отсутствуют.

Коммерческая версия Power BI Pro стоит 9,99$ за одного пользователя в месяц. Лимиты поднимаются до 10 Гб на человека и скоростью в 1 млн строк/час. Включаются все несколько дополнительных функций, например каталог данных и продвинутые инструменты совместной работы.

Плюсы и минусы Power BI показаны на рисунке 5:



Рисунок 5 – плюсы и минусы Power BI

1.3.3 Tableau

Tableau специализируется на анализе данных через их визуализацию. В нем легко создавать интерактивные дашборды, которые позволяют изучить динамику, тренды и структуру данных, с помощью удобных и простых, но не менее эффективных графиков [22].

Средства доступные для подключения к Tableau изображены на рисунке 6:



Рисунок 6 – средства доступные для подключения к Tableau

Особенность Tableau заключается в комбинировании данных из разных баз данных и источников, вкупе с высоким многообразием визуализации, позволяющей достигнуть высокого уровня понимания данных [23].

Одновременная работа нескольких пользователей над отчётом в реальном времени – удобная функция, позволяющая повысить эффективность работы.

Делиться отчётами можно с помощью трёх способов: по ссылке, через встроенный e-mail Tablue Reader или же публикуя на сервере этого приложения.

Данная платформа легка в использовании за счёт интуитивного восприятия интерфейса и многочисленным видеоурокам выкладываемых на просторах сети интернет.

За счёт интеграций с: R, Python и MATLAB, возможности Tableau возрастают многократно, устраняя недостатки данной платформы.

В связи с недавними изменениями приобрести Tableau Desktop для себя не выйдет, однако теперь можно приобрести подписку Tableau Creator, которая включает Tableau Desktop, Tableau Prep и лицензию Creator на Tableau Online или Tableau Server. Для физических лиц цена подписки составляет 70 долларов в месяц, оплачивается ежегодно (840 долларов). Виды подписок для команд и организаций можно наблюдать на рисунке 7:

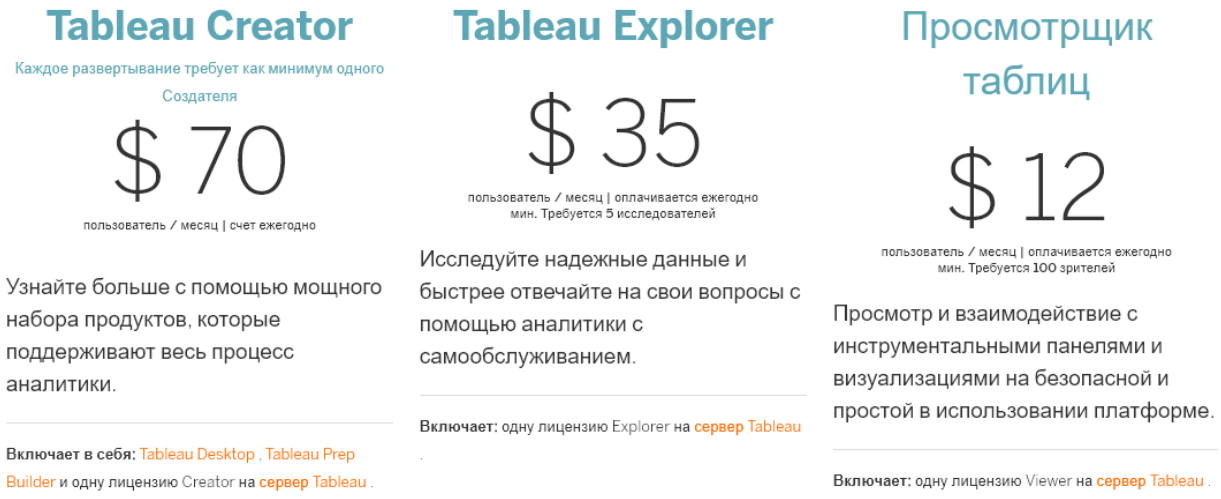

Рисунок 7 – виды подписок для команд и организаций

Плюсы и минусы Tableau отображены на рисунке 8:

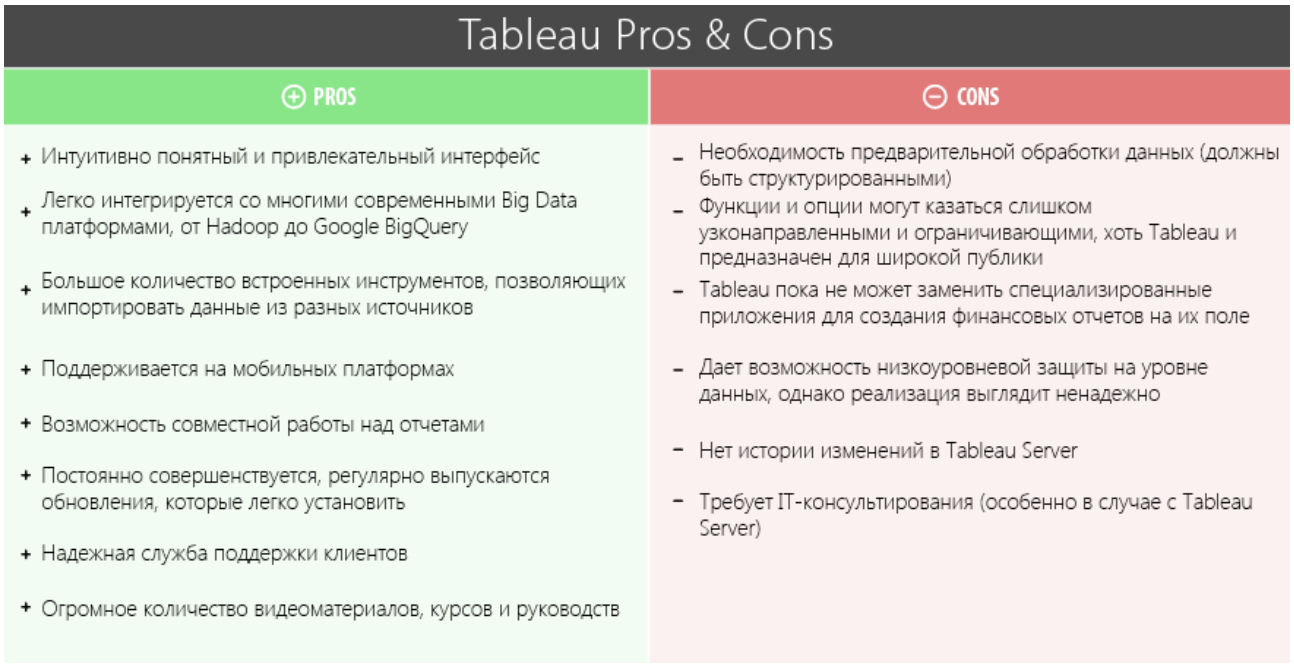

Рисунок 8 – плюсы и минусы Tableau

2 ПРИМЕНЕНИЕ НЕИЕРАРХИЧЕСКОГО МЕТОДА КЛАСТЕРИЗАЦИИ И МЕТОДОВ ВИЗУАЛИЗАЦИИ ДЛЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ СРЕДСТВАМИ TABLEAU

2.1 СУЩНОСТЬ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ И МЕТОДОВ ВИЗУАЛИЗАЦИИ В АНАЛИЗЕ ДАННЫХ

Кластеризация является одной из фундаментальных задач в области анализа данных и часто выступает ее первым шагом. На практике легче выделить группы схожих объектов (кластеров), изучить их особенности и построить для каждой группы отдельную модель. Таким приемом часто пользуются в маркетинге, выделяя группы клиентов, поставщиков, товаров и разрабатывая для каждой их них отдельную стратегию.

Проблемы кластеризации:

Все атрибуты, или признаки объектов делятся на числовые и категорийные. Например, атрибут "цена" – числовой, "упаковка товара" – категорийный. Большинство алгоритмов кластеризации предполагают сравнение объектов между собой на основе некоторой меры близости (сходства). Мерой близости называется величина, возрастающая с увеличением близости объектов. Меры сходства создаются по специальным правилам, а выбор конкретных мер зависит от задачи, а также от шкалы измерений. В качестве меры близости для числовых атрибутов часто используется евклидово расстояние, вычисляемое по формуле изображенной на Рисунок 9:

$$D(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

Рисунок 9 – формула евклидова расстояния

Для категорийных атрибутов распространена мера сходства Чекановского-Серенсена и Жаккара []. Мера показана на Рисунок 10:



Рисунок 10 – мера сходства Чекановского-Серенсена и Жаккара

Алгоритмы кластеризации:

По способу разбиения на кластеры алгоритмы бывают двух типов: иерархические и неиерархические. Классические иерархические алгоритмы работают с категорийными атрибутами, когда строится полное дерево вложенных кластеров. Здесь распространены агломеративные методы построения иерархий кластеров – в них производится последовательное объединение исходных объектов и соответствующее уменьшение числа кластеров. Иерархические алгоритмы обеспечивают сравнительно высокое качество кластеризации и не требуют предварительного задания количества кластеров. Неиерархические алгоритмы основаны на оптимизации целевой функции, определяющей оптимальное разбиение множества объектов на кластеры. В этой группе популярны алгоритмы семейства k-средних (k-means, fuzzy c-means, Густафсон-Кесселя и др.), которые в качестве целевой функции используют сумму квадратов взвешенных отклонений координат объектов от центров кластеров. Кластеры ищутся сферической либо эллипсоидной формы. В канонической реализации минимизация функции производится на основе метода множителей Лагранжа и позволяет найти только ближайший локальный минимум. Использование методов глобального поиска (генетические алгоритмы) значительно увеличит вычислительную сложность алгоритма. Недостатками алгоритма являются зависимость результата от инициализации центров кластеров и неопределенность выбора начального количества точек, играющих роль центров кластеров при первой итерации [6].

Среди неиерархических алгоритмов, не основанных на расстоянии, следует выделить EM-алгоритм (Expectation-Maximization). В нем вместо центров кластеров предполагается наличие функции плотности вероятности для каждого кластера с соответствующим значением математического ожидания и дисперсией. В смеси распределений ведется поиск их параметров (средние и стандартные отклонения) по принципу максимума правдоподобия. Проблема

заключается в том, что перед стартом алгоритма выдвигается гипотеза о виде распределений, которые оценить в общей совокупности данных сложно.

Еще одна проблема появляется тогда, когда атрибуты объекта смешанные – одна часть имеет числовой тип, а другая часть – категорийный. При этом категорийные атрибуты часто важнее числовых.

Алгоритм оптимизации целевой функции в неиерархических алгоритмах, основанных на расстояниях, носит итеративный характер, и на каждой итерации требуется рассчитывать матрицу расстояний между объектами. При большом числе объектов это требует значительных вычислительных ресурсов.

Алгоритмы семейства k-means хорошо работают, когда данные в многомерном пространстве образуют компактные сгустки, хорошо отличимые друг от друга. Если данные имеют вложенную форму, то алгоритм семейства k-means не справится с такой задачей. Также алгоритм плохо работает в случае, когда один кластер значительно больше остальных, и они находятся близко друг от друга – возникает эффект "расщепления" большого кластера.

Разработаны интересные расширения алгоритма k-means для работы с категорийными атрибутами (k-modes) и смешанными атрибутами (k-prototypes). Например, в k-prototypes расчет расстояний между объектами осуществляется по-разному в зависимости от типа атрибута.

Разработаны масштабируемые аналоги k-means и EM (scalable k-means и scalable EM), масштабируемые агломеративные методы (CURE, CACTUS). Эти современные алгоритмы требуют всего несколько (от двух до десяти) сканирований базы данных до получения финальной кластеризации.

Получение масштабируемых алгоритмов основано на идее отказа от локальной функции оптимизации. Парное сравнение объектов между собой в алгоритме k-means есть не что иное, как локальная оптимизация, т.к. на каждой итерации необходимо рассчитывать расстояние от центра кластера до каждого объекта. Это ведет к большим вычислительным затратам. При задании глобальной функции оптимизации добавление новой точки в кластер не требует больших вычислений: оно рассчитывается на основе старого значения, нового

объекта и так называемых кластерных характеристик (clusters features). Конкретные кластерные характеристики зависят от того или иного алгоритма. Так появились алгоритмы BIRCH, LargeItem, CLOPE и многие другие.

Перейдём к визуальной аналитике.

Основная идея визуального анализа – представление данных в различных визуальных формах, помогающих глубже понимать анализируемые данных и работать с визуальным представлением.

Раньше отображение визуальных данных в реальном времени не могло позволить себе более сложные графические решения, чем простую трехмерную или двухмерную графику из-за сложности и дороговизны. Благодаря совершенствованию операционной системы и аппаратной части, более продвинутые средства визуального анализа стали доступны. На сегодняшний день существует много различных видов графических образов, которые позволяют представлять результаты анализа в виде, удобном для понимания человеком.

Благодаря совершенствованию технологий стали доступны оценки расстояния и размеров объекта, в реальном времени рассматривать исследуемые данные со всех сторон, что открывает использования для анализа перцепционные навыки в обнаружении неопределенных образцов в визуальном пространстве. Это даёт возможность более чётко сформулировать цели исследования, когда о первичных данных мало известно, а цели непонятны.

Работа с зашумленными и неоднородными данными, с помощью визуальных данных и участия человека, легка и более точка в сравнении с автоматическими методами. Также не требуется статистических и математических алгоритмов, поскольку отображённые графические данные интуитивно понятны.

Все перечисленные преимущества позволяют еще больше облегчить работу аналитика и повысить качество получаемых знаний при совместном использовании визуального анализа и методов автоматического анализа.

Визуальный анализ данных обычно выполняется в три этапа:

- беглый анализ — позволяет идентифицировать интересные шаблоны и сфокусироваться на одном или нескольких из них;
- увеличение и фильтрация — идентифицированные на предыдущем этапе шаблоны отфильтровываются и рассматриваются в большем масштабе;
- детализация по необходимости — если пользователю нужно получить дополнительную информацию, он может визуализировать более детальные данные.

Процесс визуализации изображен на Рисунок 11. Так же, как и при анализе данных, информация извлекается из некоторого источника, например, из базы данных или из файлов. Затем к ней могут быть применены алгоритмы Data Mining для выявления скрытых закономерностей (классификации, кластеризации и т. п.). Как результаты применения алгоритмов, так и исходные данные подвергаются обработке в ядре визуализации. Основной целью обработки является приведение многомерных данных к такому виду, который можно было бы представить на экране монитора.



Рисунок 11 – процесс визуализации данных

Характеристики средств визуализации данных

Наборы визуализируемых данных, как и в Data Mining, представляют собой матрицы, в которых ряды являются данными (например, записями об экспериментах, покупки в магазине и т. п.), а колонки — атрибутами данных. При этом данные могут характеризоваться одним или несколькими атрибутами.

Кроме того, сами данные могут иметь более сложную структуру: иерархическую, текстовую, графическую и т. п. Таким образом, выделяют следующие виды данных, с которыми могут работать средства визуализации:

- одномерные данные — одномерные массивы, временные ряды и т. п.;
- двумерные данные — точки двумерных графиков, географические координаты и т. п.;
- многомерные данные — финансовые показатели, результаты экспериментов и т. п.;
- тексты и гипертексты — газетные статьи, Web-документы и т. п.;
- иерархические и связанные — структура подчиненности в организации, электронная переписка людей, гиперссылки документов и т. п.;
- алгоритмы и программы — информационные потоки, отладочные операции и т. п.

Одномерные данные состоят из одного атрибута. Пример: временные ряды, атрибут – время. Возможно, что несколько значений будут связаны с одной отметкой времени.

Двумерные данные – два отдельных измерения. Типичным примером могут служить географические данные, характеризующиеся двумя атрибутами: долгота и широта. Обычно такие данные отображаются в виде точек в двумерной системе координат.

Этим дум видам следует опасаться большого объема данных, поскольку возможны потери при попытке уменьшения масштаба для корректного отображения данных.

Многомерные данные. Отображение в виде двух/трехмерных точек невозможно. Например, реляционные таблицы баз данных. Атрибуты – колонки данных. Для корректного отображения требуются специальные методы визуализации, например, параллельные координаты.

Тексты и гипертексты – не могут быть отражены в виде измерений и не все методы визуализации для них могут использоваться, поэтому необходимо подготовить данные к виду пригодному для методов визуализации.

Иерархические и связанные данные. Для представления таких взаимосвязей широко используются графы. Граф состоит из набора узлов и соединяющих их линий, называемых дугами. Примерами таких данных могут являться: электронная почта, пересылаемая между людьми, гиперссылки между Web-документами, файловая структура диска и т. п. Для отображения таких связей используются специальные методы визуализации.

Другой класс данных — алгоритмы и программы. Копирование больших программных проектов является проблемой. Целью визуализации является поддержка в разработке программных средств, например, отображение потока информации в программе для лучшего понимания алгоритмов и написанного кода, отображение структуры тысячи строк исходного кода в виде графа и поддержки процесса отладки, визуализация ошибок. Существует большое количество инструментов и систем, которые реализуют эти задачи.

Для визуализации перечисленных типов данных используются различные визуальные образы и методы их создания. Очевидно, что количество визуальных образов, которыми могут представляться данные, ограничиваются только человеческой фантазией. Основное требование к ним — это наглядность и удобство анализа данных, которые они представляют. Методы визуализации могут быть как самые простые (линейные графики, диаграммы, гистограммы и т. п.), так и более сложные, основанные на сложном математическом аппарате. Кроме того, при визуализации могут использоваться комбинации различных методов. Выделяют следующие типы методов визуализации:

- стандартные 2D/3D-образы — гистограммы, линейные графики и т. п.;
- геометрические преобразования — диаграмма разброса данных, параллельные координаты и т. п.;
- отображение иконок — линейчатые фигуры (needle icons) и звезды (star icons);
- методы, ориентированные на пикселы — рекурсивные шаблоны, циклические сегменты и т. п.;

- иерархические образы — древовидные карты и наложение измерений [2, стр. 193].

Простейшие методы визуализации, к которым относятся 2D/3D-образы, широко используются в существующих системах (например, в Microsoft Excel). К этим методам относятся: графики, диаграммы, гистограммы и т. п. Основным их недостатком является невозможность приемлемой визуализации сложных данных и большого количества данных.

Методы геометрических преобразований визуальных образов направлены на трансформацию многомерных наборов данных с целью отображения их в декартовом и в недекартовом геометрических пространствах. Данный класс методов включает в себя математический аппарат статистики. К нему относятся такие популярные методы, как диаграмма разброса данных (scatter plot), параллельные координаты (Parallel Coordinates), гипердоли (Hyperslice) и др.

Другим классом методов визуализации данных являются методы отображения иконок. Их основной идеей является отображение значений элементов многомерных данных в свойства образов. Такие образы могут представлять собой: человеческие лица, стрелки, звезды и т. п. Визуализация генерируется отображением атрибутов элементов данных в свойства образов. Такие образы можно группировать для целостного анализа данных. Результирующая визуализация представляет собой шаблоны текстур, которые имеют различия, соответствующие характеристикам данных и обнаруживаемые преаттентивным восприятием.

Основной идеей методов, ориентированных на пикселы, является отображение каждого измерения значения в цветной пиксел и их группировка по принадлежности к измерению. Так как один пиксел используется для отображения одного значения, то, следовательно, данный метод позволяет визуализировать большое количество данных (свыше одного миллиона значений).

Методы иерархических образов предназначены для представления данных, имеющих иерархическую структуру. В случае многомерных данных

должны быть правильно выбраны измерения, которые используются для построения иерархии.

В результате применения методов визуализации будут построены визуальные образы, отражающие данные. Однако этого не всегда бывает достаточно для полного анализа. Пользователь должен иметь возможность работать с образами: видеть их с разных сторон, в разном масштабе и т. п. Для этого у него должны быть соответствующие возможности взаимодействия с образами:

- динамическое проецирование;
- интерактивная фильтрация;
- масштабирование образов;
- интерактивное искажение;
- интерактивное комбинирование.

Основная идея динамического проецирования заключается в динамическом изменении проекций при проведении исследования многомерных наборов данных. Примером может служить проецирование в двумерную плоскость всех интересующих проекций многомерных данных в виде диаграмм разброса (scatter plots). Необходимо обратить внимание, что количество возможных проекций экспоненциально увеличивается с ростом числа измерений, и следовательно, при большом количестве измерений проекции будут тяжело воспринимаемы.

При исследовании большого количества данных важно иметь возможность разделять наборы данных и выделять интересующие поднаборы — фильтровать образы. При этом важно, чтобы данная возможность предоставлялась в режиме реального времени работы с визуальными образами (т. е. интерактивно). Выбор поднабора может осуществляться или напрямую из списка, или с помощью определения свойств интересующего поднабора. Выбор из списка неудобен при большом количестве поднаборов, в то же время запросы не всегда позволяют получить желаемый результат.

Примером масштабирования образов является "магическая линза" (Magic Lenses). Ее основная идея состоит в использовании инструмента, похожего на увеличительное стекло, чтобы выполнять фильтрацию непосредственно при визуализации. Данные, попадающие под увеличительное стекло, обрабатываются фильтром, и результат отображается отдельно от основных данных. Линза показывает модифицированное изображение выбранного региона, тогда как остальные визуализированные данные не детализируются.

Масштабирование — это хорошо известный метод взаимодействия, используемый во многих приложениях. При работе с большим объемом данных этот метод хорош для представления данных в сжатом общем виде, и, в то же время, он предоставляет возможность отображения любой их части в более детальном виде. Масштабирование может заключаться не только в простом увеличении объектов, но в изменении их представления на разных уровнях. Так, например, на нижнем уровне объект может быть представлен пикселом, на более высоком уровне — неким визуальным образом, а на следующем — текстовой меткой.

Метод интерактивного искажения поддерживает процесс исследования данных с помощью искажения масштаба данных при частичной детализации. Основная идея этого метода заключается в том, что часть данных отображается с высокой степенью детализации, а одновременно с этим остальные данные показываются с низким уровнем детализации. Наиболее популярные методы — это гиперболическое и сферическое искажения, которые часто используются на иерархиях и графах, но могут применяться и в других визуальных образах.

Существует достаточно много методов визуализации, но все они имеют как достоинства, так и недостатки. Основная идея комбинирования заключается в объединении различных методов визуализации для преодоления недостатков одного из них. Различные проекции рассеивания точек, например, могут быть скомбинированы с методами окрашивания и компоновки точек во всех проекциях. Такой подход может быть использован для любых методов визуализации. Окраска точек во всех методах визуализации дает возможность

определить зависимости и корреляции в данных. Таким образом, комбинирование нескольких методов визуализации обеспечивает большую информативность, чем в общем независимое использование методов. Типичными примерами визуальных образов, которые могут комбинироваться, являются: точки рассеивания, гистограммы, параллельные координаты, отображаемые пикселы и карты.

Любое средство визуализации может быть классифицировано по всем трем параметрам, т. е. по виду данных, с которым оно работает, по визуальным образам, которые оно может предоставлять, и по возможностям взаимодействия с этими визуальными образами. Очевидно, что одно средство визуализации может поддерживать разные виды данных, разные визуальные образы и разные способы взаимодействия с образами.

2.2 ПРОЦЕСС КЛАСТЕРИЗАЦИИ В СРЕДЕ TABLEAU

Кластерный анализ разбивает точки на диаграмме рассеяния на кластеры, в которых точки из одного кластера сильнее похожи друг на друга, чем на точки других кластеров. Tableau присваивает каждому кластеру свой цвет.

Для кластерного анализа Tableau использует метод k-средних. По этому методу данные разбиваются на несколько сегментов. Для каждого сегмента подбирается центр кластера, до тех пор, пока суммарное квадратичное отклонение точек кластеров от центров этих кластеров не будет сведено к минимуму. Иными словами, метод k-средних, вместе с Tableau, пытается определить схожесть точек на основе их расстояния до центра кластера. Этот метод не всегда работает хорошо. Результат зависит от выбора исходных центров кластеров, их оптимальный выбор неизвестен.

Для того, чтобы провести кластеризацию, необходимо добавить данные на полки Columns, Rows, Filters и Marks. затем перейти на вкладку Analytics и

перетащить модель Cluster на точечную диаграмму. Пример показан на рисунках: 12, 13, 14 соответственно.



Рисунок 12 – добавление модели "Кластер" на точечную диаграмму

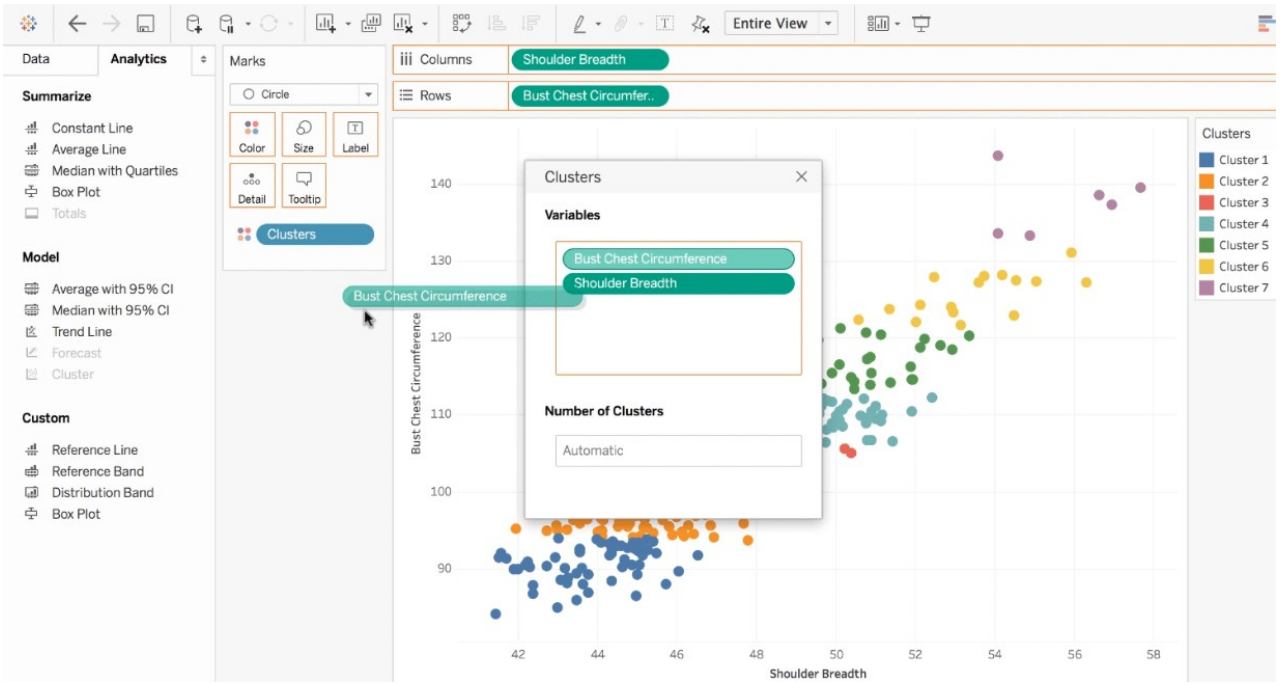

Рисунок 13 – очистка от лишних значений

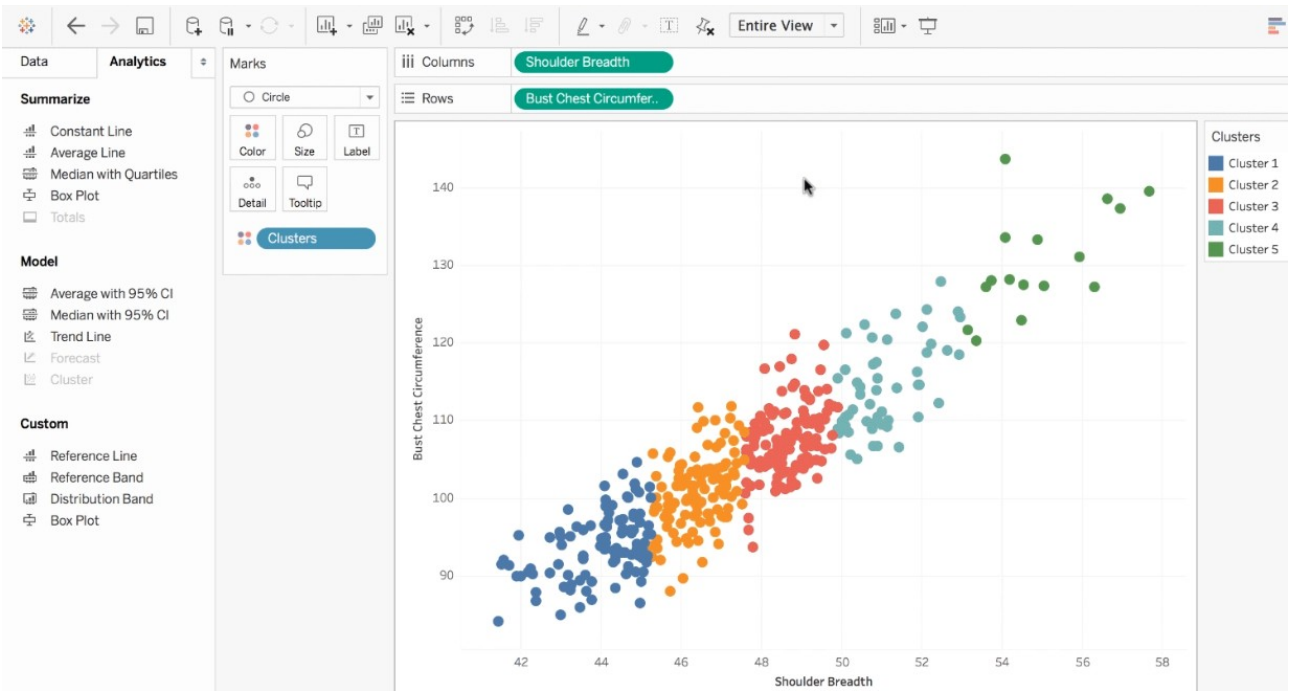

Рисунок 14 – результат кластеризации

В результате образовалось 5 кластеров.

Кластеризация – описательная процедура, которая не делает статистических выводов, однако с её помощью появляется возможность провести первичный, разведочный анализ и изучить структуру данных.

Благодаря разделению выбранных групп на кластеры возможно установить высокодоходную, средне доходную и низкодоходную выручку по товарам, выявить лидеров продаж, как по группам, так и по отдельным товарам, определить дешёвые, но популярные товары, менее популярные, но дорогие. А также выявить факторы, влияющие на вышеописанные возможности. Данный подход позволяет не только существенно сократить объекты анализа, а дополнительно индивидуально подойти к каждой группе товаров.

2.3 ПРОЦЕСС ВИЗУАЛИЗАЦИИ В СРЕДЕ TABLEAU

Поскольку Tableau – аналитическая платформа, специализирующаяся на средствах визуализации данных, большинство действия с данными, помимо подключения источников данных будут считаться визуализацией.

Проиллюстрируем несколько методов визуализации с помощью среды Tableau.

2D/3D

Добавляем на полку Columns дату, указываем сортировку по месяцам, на полку Rows перетащим выручку, в Filters добавим группу и выберем сортировку, например «Энергетические товары» и выберем тип представления данных lines (Рисунок 15).

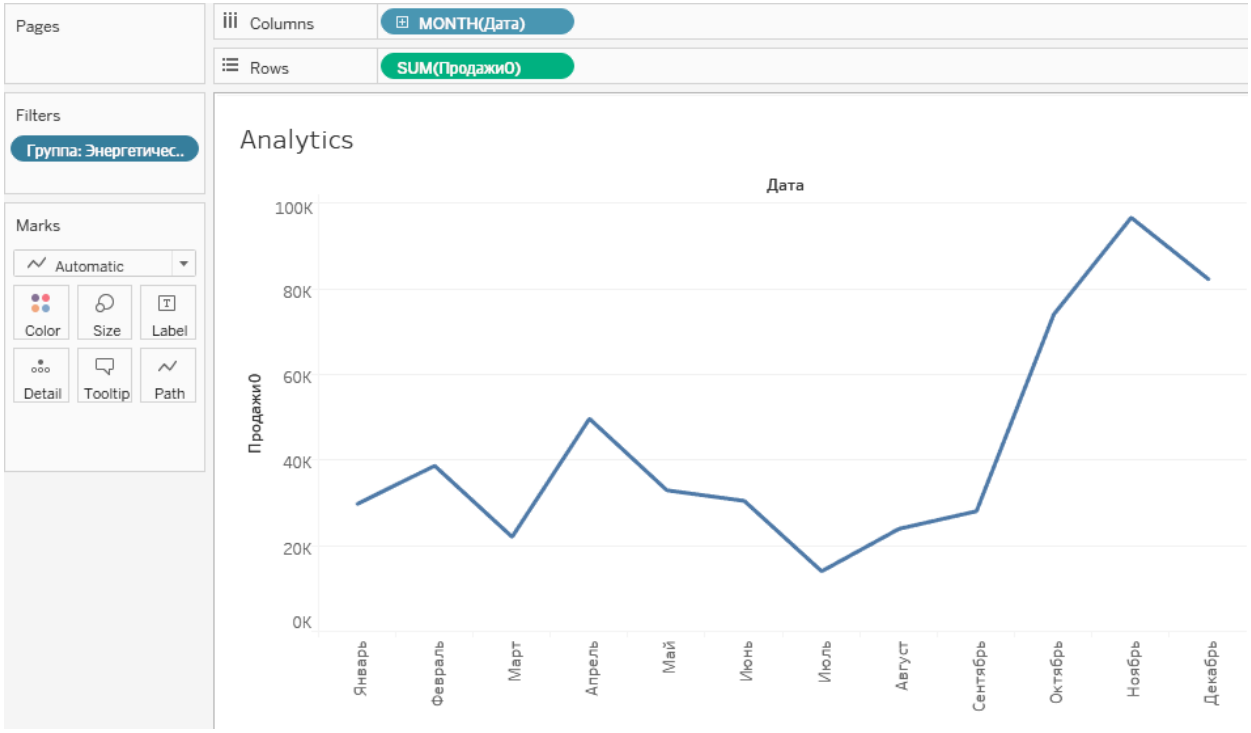

Рисунок 15 – линейный 2D график продаж в разрезе месяцы

В результате получился линейный график продаж в разрезе месяцы, отображающий выручку товарной группы «Энергетические напитки».

Геометрические преобразования — диаграмма разброса данных, параллельные координаты и т. п.;

На полку Columns и Filter добавим Measure Names (Измеряемые имена), на полку Rows и Marks перетащим Measure Values (Измеряемые значения), дополнительно указав продажи, коэффициент прибыли и клиента в измеряемых значениях Marks (Рисунок 16).

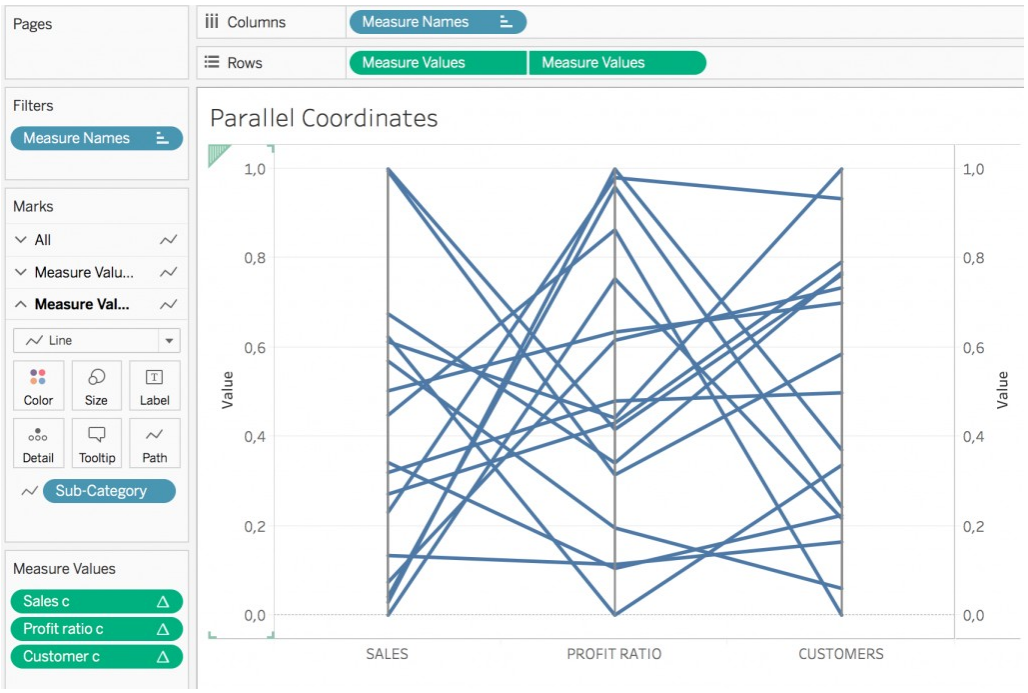

Рисунок 16 – диаграмма параллельных координат

В итоге получили диаграмму параллельных координат, с помощью которой можно выявить паттерны и корреляцию между разными переменными.

Иерархические образы

Добавим в filters Группы, в Marks товар, количество, и выручку. Выберем тип визуализации treemaps (Рисунок 17).

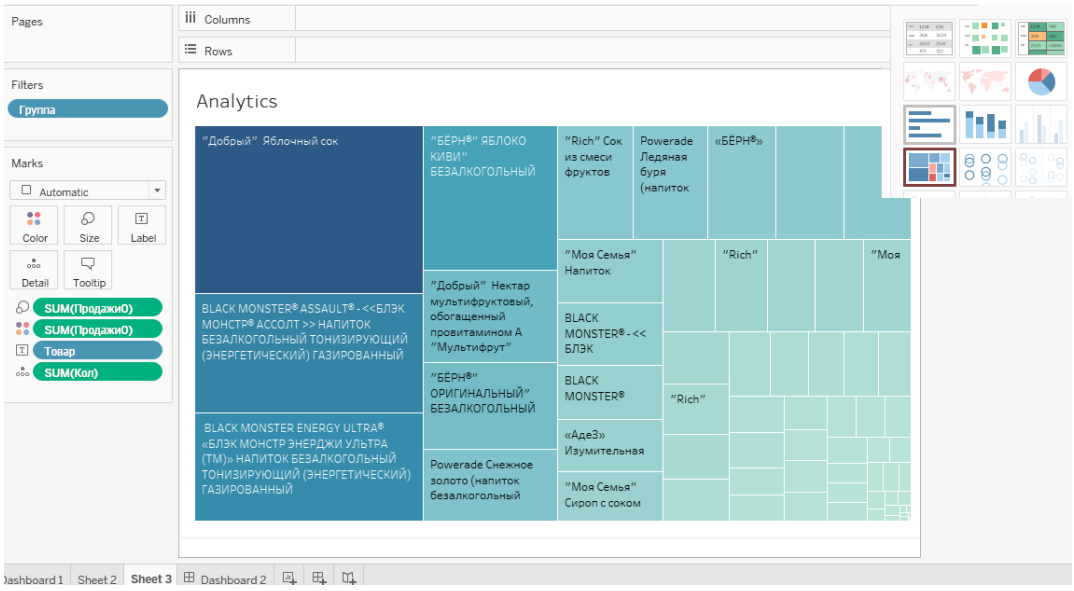

Рисунок 17 – график древовидная карта

В итоге получили древовидную карту, отображающую иерархию товаров по объёму выручки относительно других товаров.

В данном параграфе были рассмотрены способы реализации некоторых методов визуализации на примере Tableau.

3 ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ПРОДАЖ В СРЕДЕ TABLEAU

3.1 ПОДКЛЮЧЕНИЕ ИСТОЧНИКА ДАННЫХ

Запускаем программу Tableau 2020.1, (Рисунок 18).



Рисунок 18 – Tableau Desktop версии 2020.1

В стартовом окне выбираем соединитель Text file, показанный на Рисунок 19.

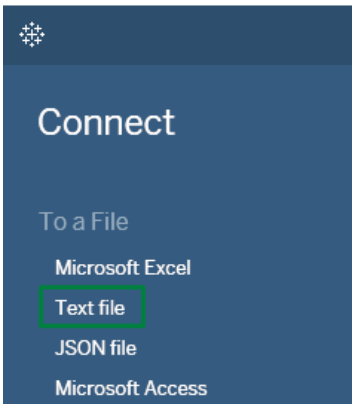

Рисунок 19 – выбор средства для импорта данных

Выбираем данные для импорта, (Рисунок 20).

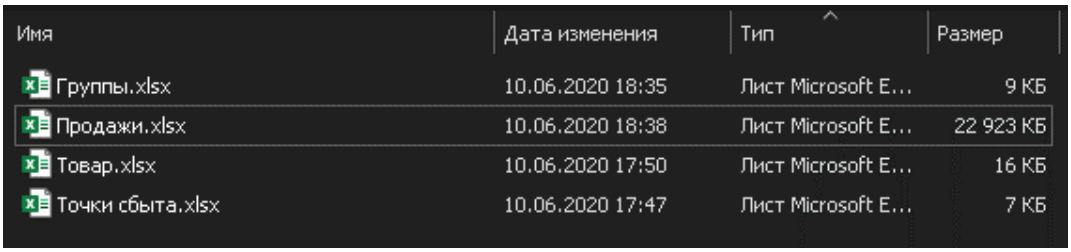

Рисунок 20 – данные для импорта

В контекстном меню файла «Продажи» переключаемся на «Dield names are in first row», означающий «Имя поля в первой строке», (Рисунок 21).

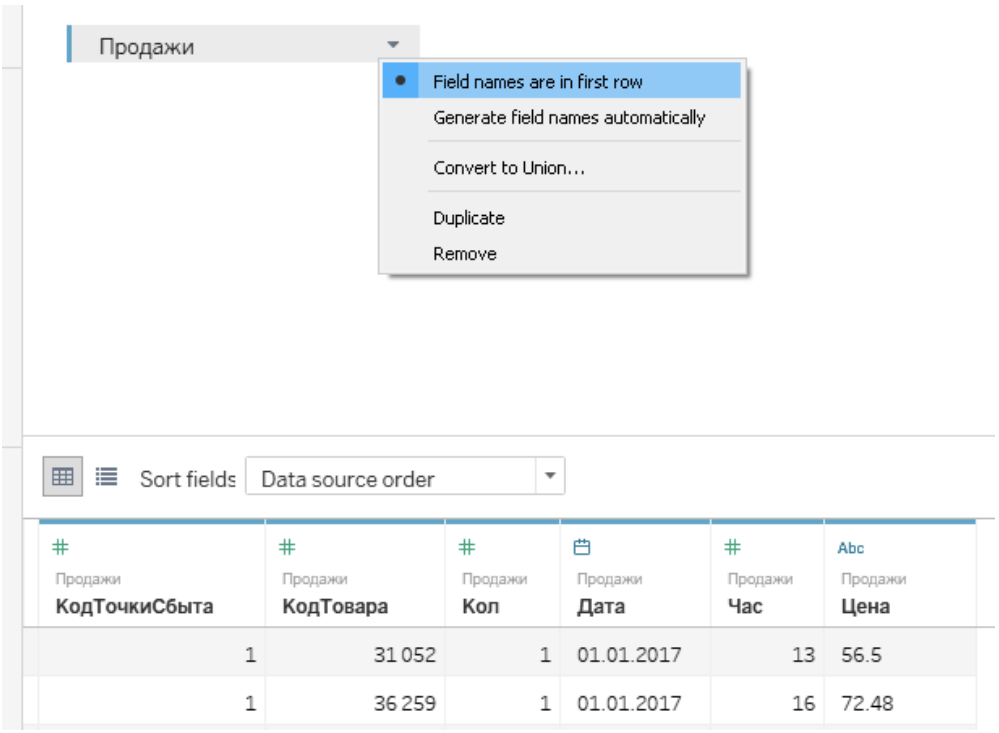

Рисунок 21 – обозначение имя поля для первой строки

Связываем две таблицы «Продажи» и «Товар» по полю «КодТовара» с помощью типа объединения «Inner Join». Связь показана на Рисунок 22.

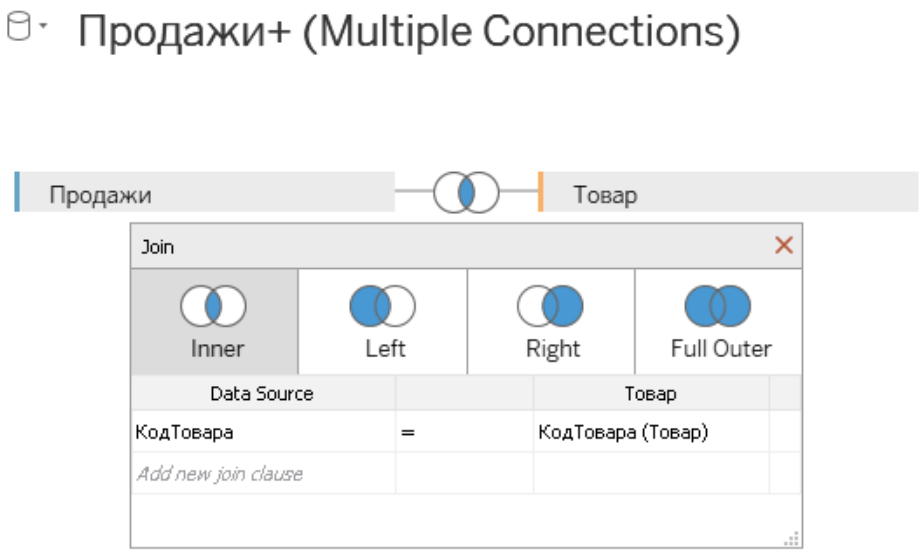

Рисунок 22 – установка типа связи для двух таблиц

Подключаем таблицу «Точки сбыта», связываем с таблицей «Продажи» по полю «КодТочкиСбыта», тип объединения – Inner Join. Связь показана на Рисунок 23.

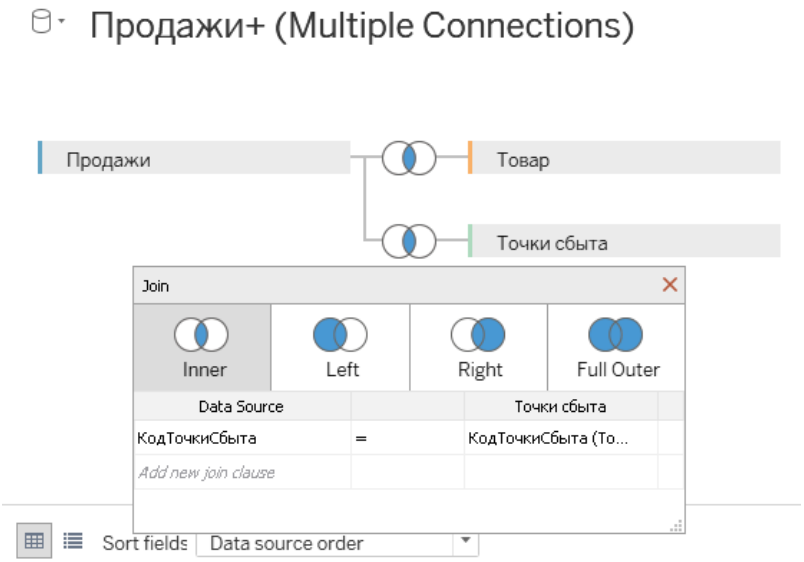

Рисунок 23 – установка типа связи для двух таблиц

Подключаем таблицу «Группы» к таблице «Товар» по полю «КодГруппы», тип объединения – Inner Join. Связь показана на Рисунок 24.

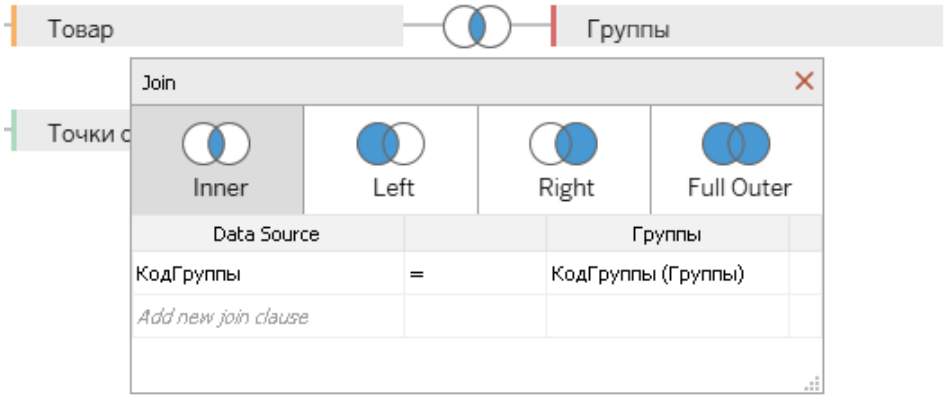

Рисунок 24 – установка типа связи для двух таблиц

Все таблицы связанные таблицы отображены на Рисунок 25.

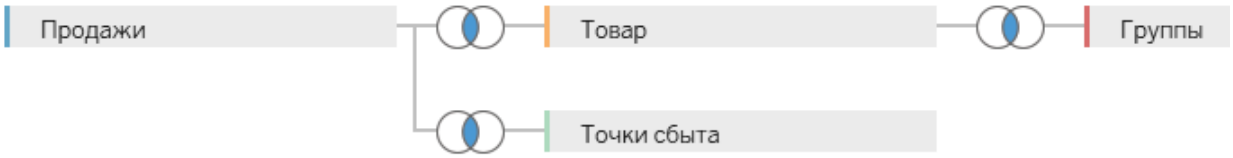

Рисунок 25 – все связанные таблицы

Получившиеся объединенные данные представлены на Рисунок 26.

Группы Группа	Продажи КодТочкиСбыта	Продажи КодТовара	Продажи Кол	Продажи Дата	Продажи Час	Продажи Цена	Товар КодТовара (Товар)
Газированные напит...	1	31 052	1	01.01.2017	13	56.5	31 052
Соки и сокосодержа...	1	36 259	1	01.01.2017	16	72.48	36 259
Газированные напит...	1	40 345	1	01.01.2017	15	15.84	40 345
Газированные напит...	1	65 841	1	01.01.2017	14	163.02	65 841
Соки и сокосодержа...	1	11 846	1	01.01.2017	13	173.32	11 846
Соки и сокосодержа...	1	45 806	1	01.01.2017	13	47.95	45 806

Рисунок 26 – объединенные данные

Цвет над столбцом обозначает из какой таблицы были взяты данные. Синий – «Продажи», желтый – «Товар», зелёный – «Точки сбыта», красный – «Группы».

Создадим вычисляемое поле. Первым действием нажимаем правой кнопкой мыши по свободному пространству в зоне «Dimensions», после чего нам откроется меню с возможностью выбора действия. «Create Calculated Field…» – создание вычисляемого поля. После того, как мы прописали нужное нам действие, программа выдала ошибку, которая заключается в различии форматов для вычисляемых данных. Исправить проблему можно изменив формат у параметра «Цена», со «String» на «Number (decimal). Вышеописанные действия изображены на Рисунок 27.

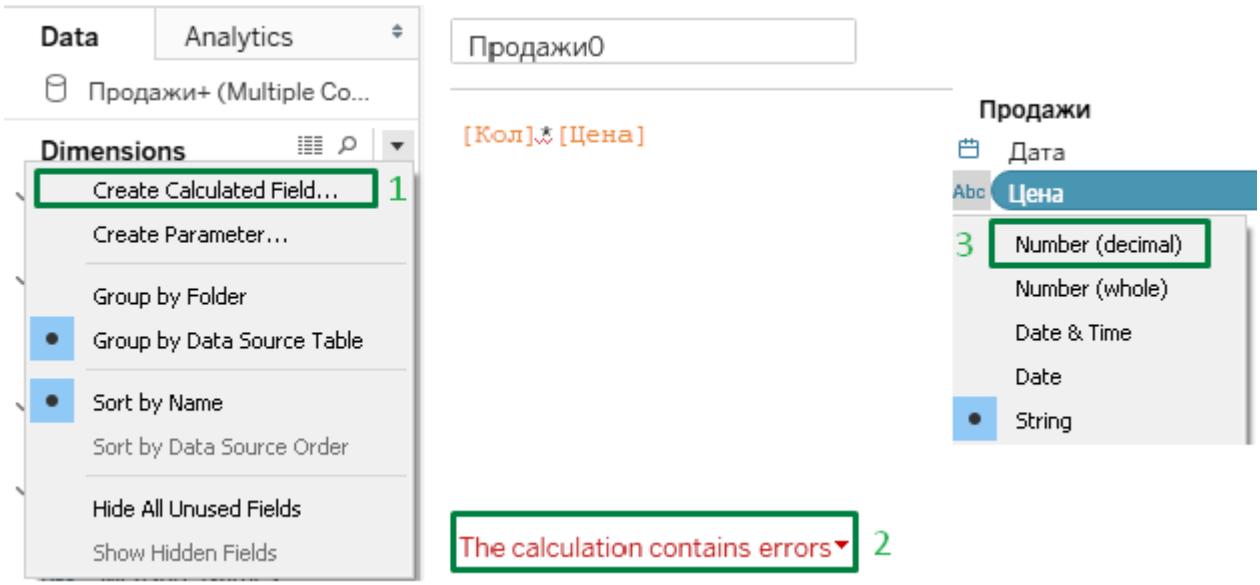

Рисунок 27 – создание вычисляемого поля

Также для дальнейшего удобства работы с данными преобразуем тип поля «Час» из числового в строковый, пример можно увидеть на Рисунок 28.

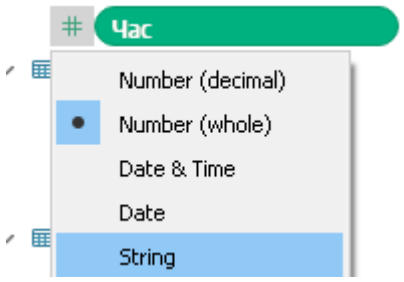

Рисунок 28 – преобразование формата в строковый

Подключение данных и их преобразование закончено.

В текущем параграфе было описано подключение, объединение, создание вычисляемого поля и изменение формата данных для последующей корректной работы с анализом данных.

Следующим этапом используем один из методов Data Mining, а именно кластерный анализ.

3.2 КЛАСТЕРНЫЙ АНАЛИЗ ПРОДАЖ

Первым пунктом создадим визуализации. Перенесём на полку «Columns» поле «Кол», на «Row» поле «Продажи», на «Filters» поле «Группа» (включая все группы), на карту «Marks» поле «Товар». В «Show Me» выберем рекомендуемый тип визуализации (Рисунок 29).

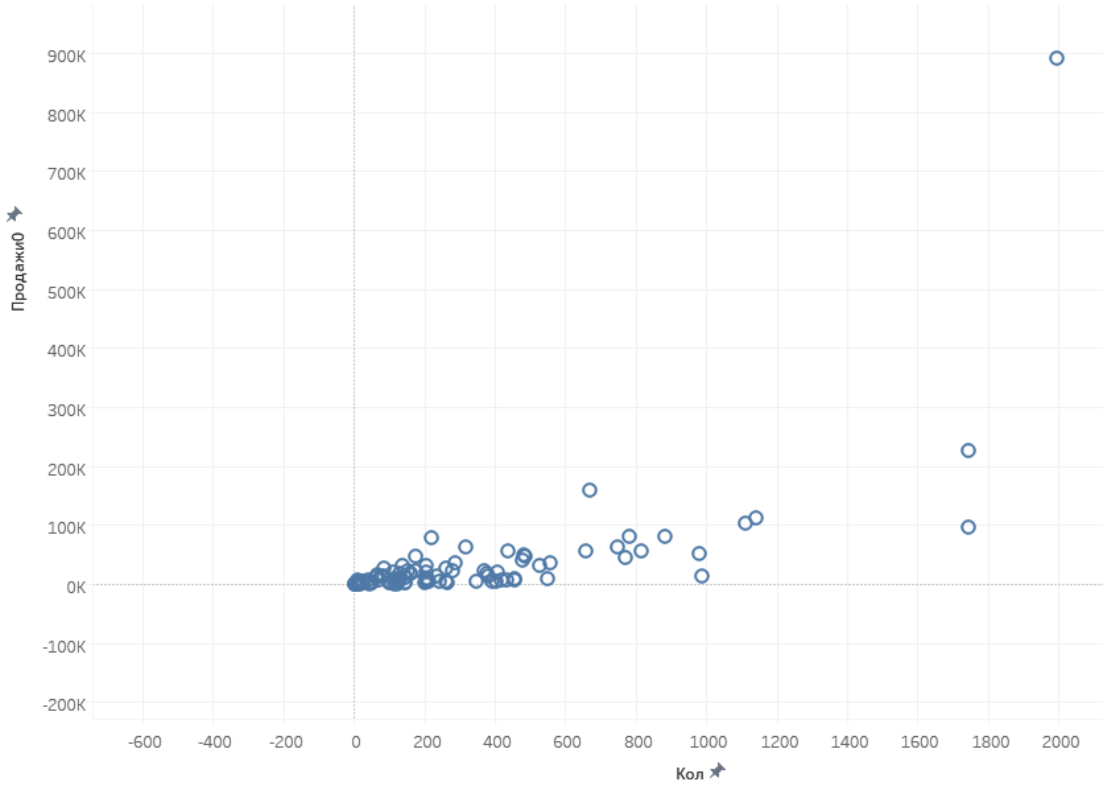

Рисунок 29 – точечная диаграмма

Переходим на вкладку «Analytics» и перенесём «Model Claster» в область построения визуализации. Укажем «Claster». Результат отображён на Рисунок 30.

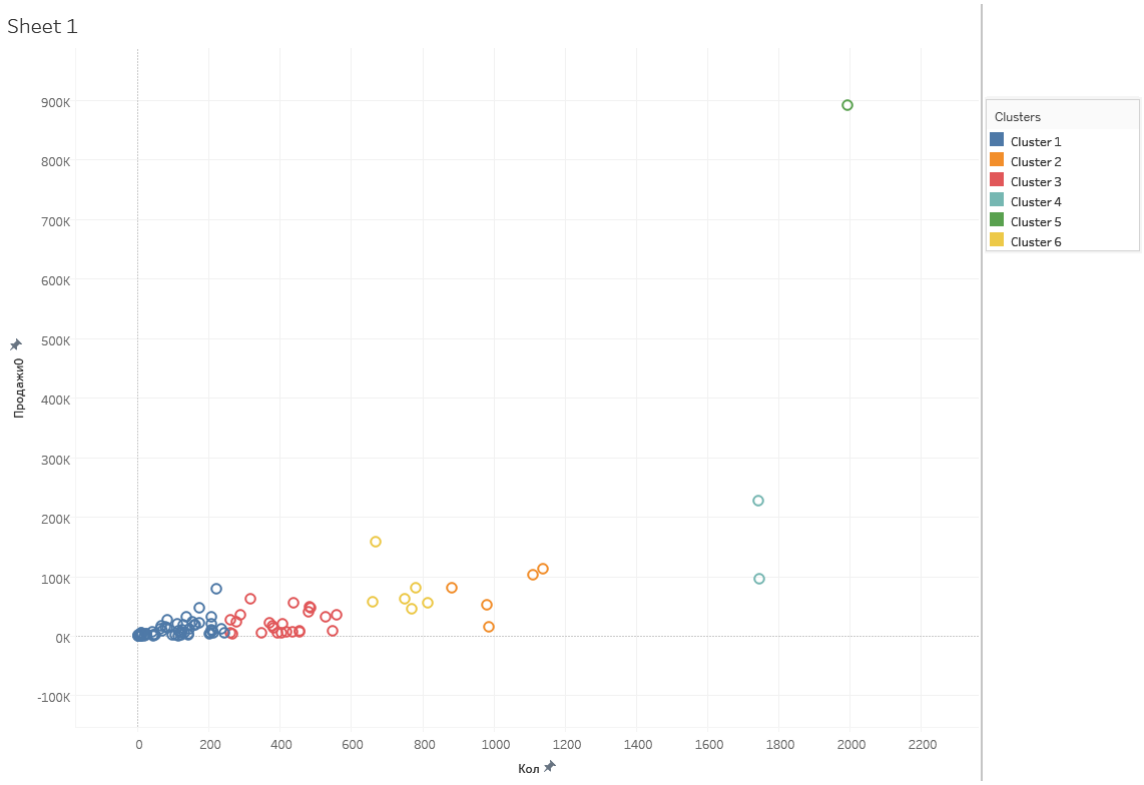

Рисунок 30 – кластеризация

Кластеризация распределила товарные группы на 6 кластеров с цветовой индикацией по умолчанию. Нас интересуют три позиции с высокими продажами в денежном выражении и высокой популярностью, посмотреть которые возможно на Рисунок 31.



Рисунок 31 – исследование высокодоходных и популярных товаров

Явным лидеров по количеству продаж, которые составляют 1995 единиц, так и по доходам, составляющих 891542 денежных единиц, является продукт «Coca-Cola напиток безалкогольный, сильногазированный», являющийся единственным в кластере 5. На втором месте находится товар «Sprite напиток безалкогольный, сильногазированный», количество продаж – 1743, выручка – 226905 д.е., принадлежащий кластеру 4, в который входит две товарные позиции. Третье место занимает товар «Добрый, яблочный сок», количество продаж – 669, выручка – 158855 д.е., входит в кластер 6, в котором находится 5 товарных позиций. Остальная номенклатура сосредоточена в первом, втором и третьем кластере.

Из четырёх товарных групп рассмотрим структуру продаж по трём их них, которые имеют наиболее высокие показатели по выручке. На первом месте «Газированные напитки», на втором «Соки и сокосодержащие напитки», на третьем «Энергетические напитки». Показатели четырёх товарных групп можно рассмотреть на Рисунок 32.

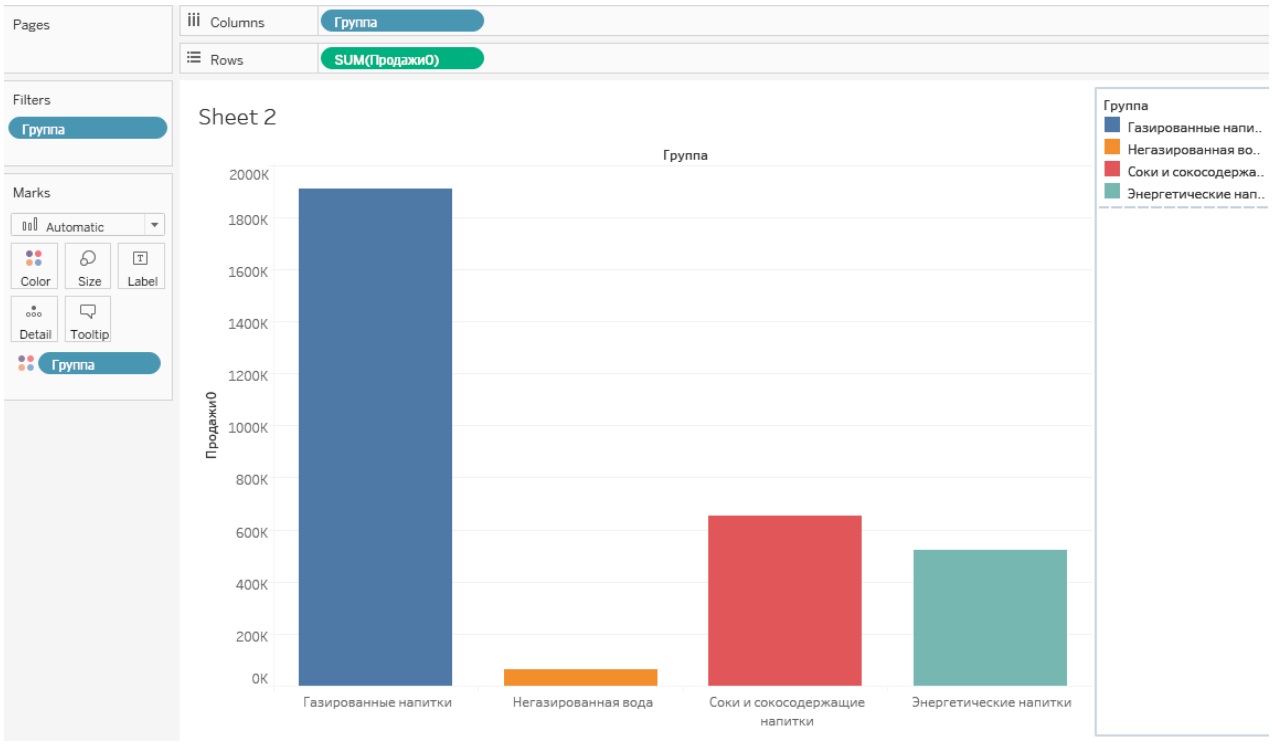

Рисунок 32 – показатели четырёх товарных групп

Рассмотрим товарную группу «Газированные напитки». В неё входят пять кластеров.

Лидером является четвёртый кластер с одной товарной позицией – «Coca-Cola напиток безалкогольный, сильногазированный». Высокая выручка, высокая популярность. Остальные кластеры имеют умеренную и низкую выручку при широкой вариации продаж в количественном выражении. Данные отображены на Рисунок 33.

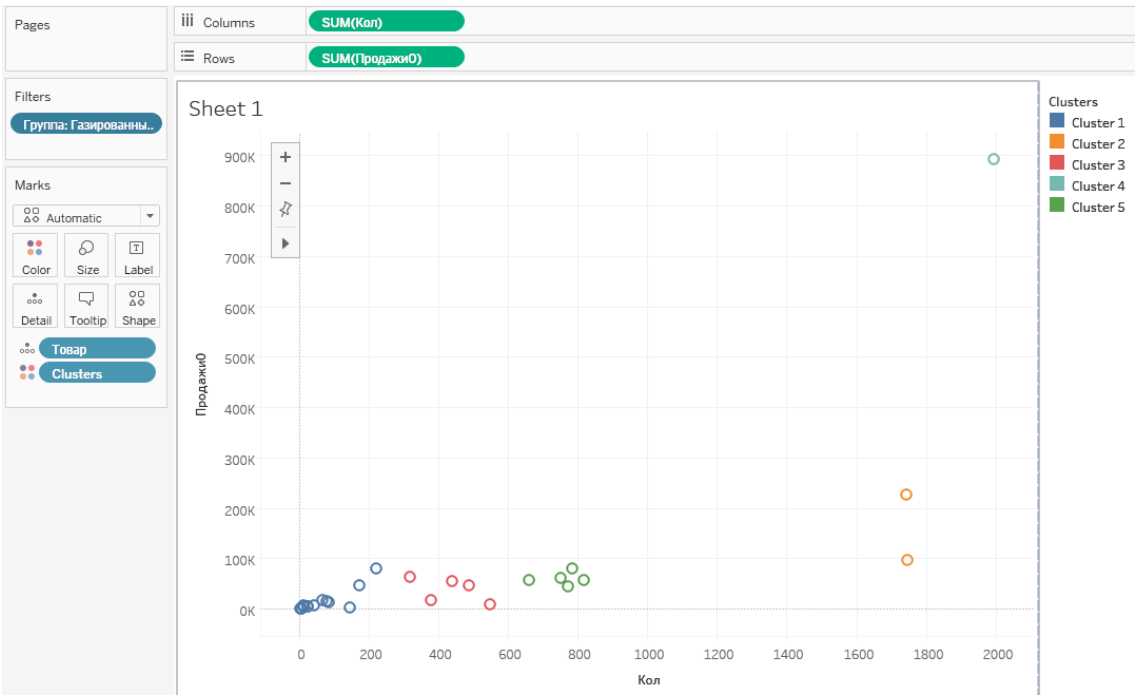

Рисунок 33 – кластеризация по группе "Газированные напитки"

Кластер 2 включает в себя два товара «Sprite, напиток безалкогольный, сильногазированный» и «Shweppes, Битер Лемон (напиток безалкогольный, сильногазированный». Характеризуется высокой популярностью, умеренной выручкой.

Кластер 3 характеризуется умеренной популярностью, небольшой выручкой.

Кластер 5 имеет 5 товарных позиций, характеризуется умеренной популярностью и продажами.

Кластер 1 имеет большое количество товарных позиций, однако показатели продаж и популярности низкие.

При формировании закупочной политики для данной товарной группы в целом по сети следует учитывать особенности продаж во всех пяти кластерах. Кроме того, необходимо учитывать динамику (сезонные особенности), взаимодействие с другими товарными группами (перекрестные продажи), вариативность продаж и другие факторы.

Рассмотрим товарную группу «Соки и сокосодержащие напитки», диаграмма представлена на Рисунок 34.

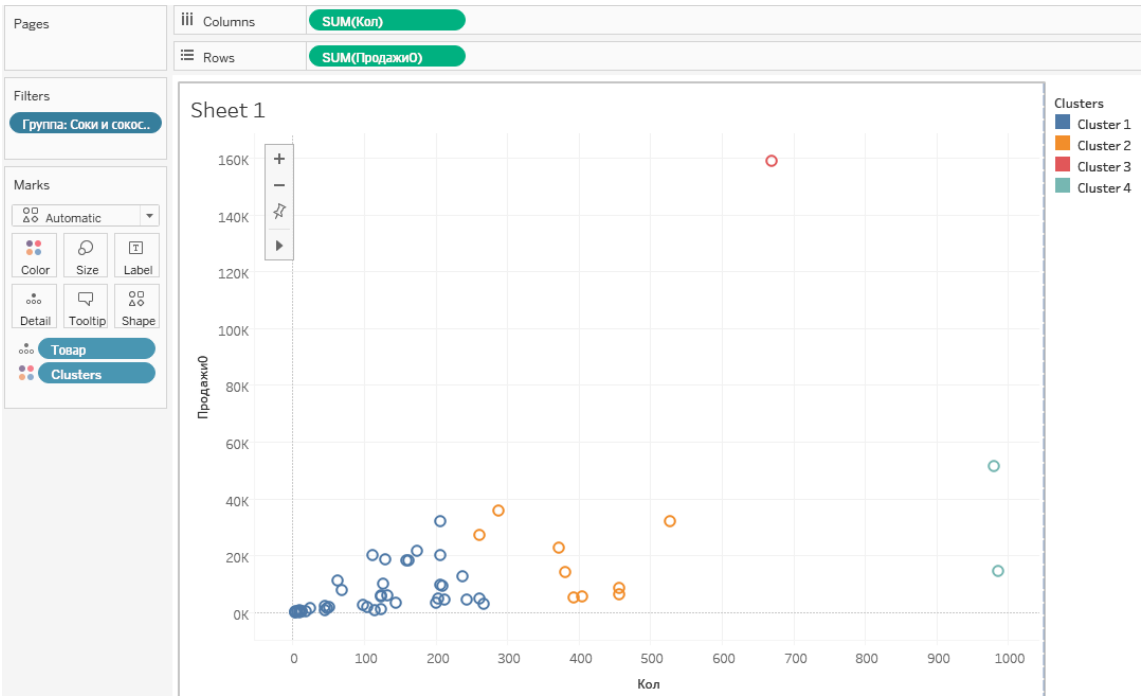

Рисунок 34 – кластеризация по группе "Соки и сокосодержащие напитки"

Наблюдаем четыре кластера. Среди данной группы безусловным лидером по продажам является «Добрый, Яблочный сок», входящий в кластер номер три.

В четвёртый кластер попали две позиции сока «Добрый» со вкусами мульти фрукт и ананас. Оба товара показывают высокие показатели популярности, однако продажи выше у мульти фруктового сока. Соку со вкусом ананаса необходимо поднять цену.

Во втором кластере 9 товаров, характеризующиеся невысокими продажами с умеренной популярностью. Выгоднее всего поднять цену тем товарам, которые имеют более высокое количество продаж, но невысокую цену.

Первый кластер имеет достаточно низкую популярность и продажи товаров в количественном выражении, однако его многочисленность приносит 39,7% от всех продаж соков в рассматриваемой товарной группе.

Рассмотрим товарную группу «Энергетические напитки», представленную на Рисунок 35.



Рисунок 35 – кластеризация по группе "Энергетические напитки"

Второй кластер, состоящий из двух товарных позиций, возглавляет список по популярности и сумме прибыли в данной товарной группе. Этот кластер должен быть под постоянным контролем с разработкой мероприятий по повышению прибыльности.

Первый кластер имеет один товар и достаточно высокую популярность и хорошей прибылью. Имеет смыл работать с данным кластером по схеме первого.

В кластере под номером четыре один товар, имеет умеренное количество продаж и выручку, такие же показатели имеет кластер номер 6, имеющий два товара. Уместным решением будет небольшое повышение цены продажи.

Второй кластер включает в себя 3 товарные позиции, характеризуется низкой популярностью и умеренными продажами.

В кластере пять всего две товарные позиции, с низкой популярностью и выручкой. Требуется более детальный анализ, для установления причины нерентабельности товаров.

Применим встроенную аналитику (embedding) для исследования продаж. Перенесем поле Цена на Columns, изменив measure с SUM на Average (средняя). Поле Выручка на Rows. Товар на Цвет. На фильтр Точка сбыта и Товар. Выведем фильтры на просмотр Show Filters для управления Viz (справа) (Рисунок 36).

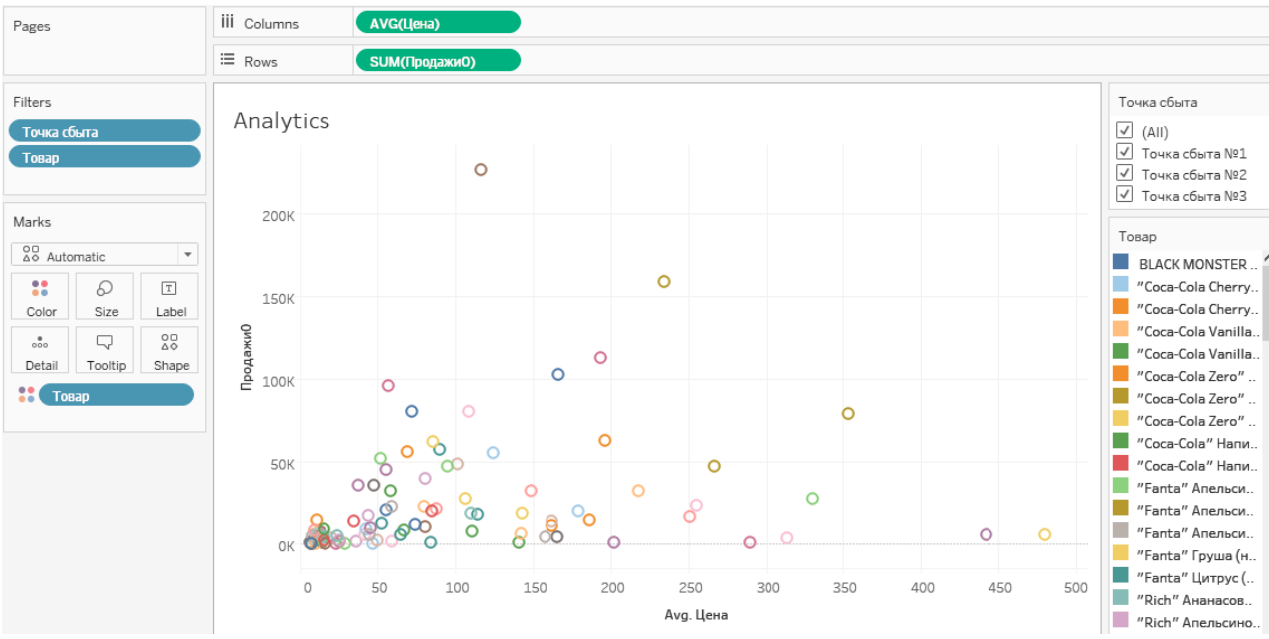

Рисунок 36 – точечная диаграмма по продажам товаров

Перенесём кластер в редактор и рассмотрим первую точку сбыта (Рисунок 37).

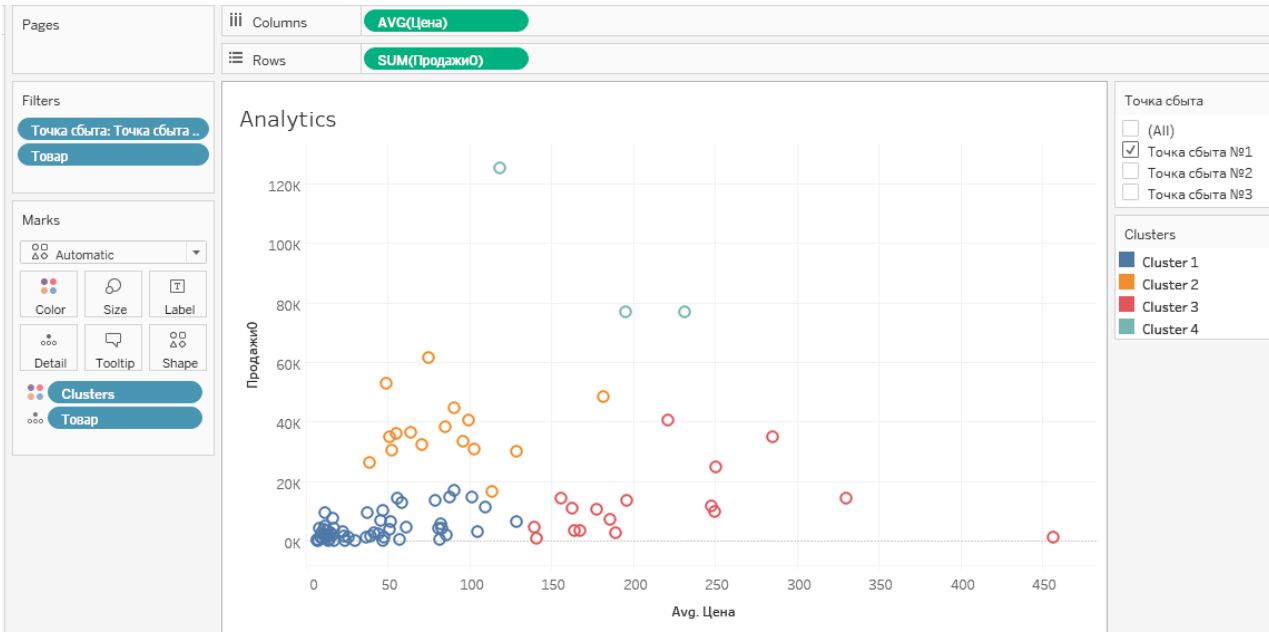

Рисунок 37 – кластеризация по первой точки сбыта

Наблюдаем 4 кластера. Создадим набор данных, покажем высокие продажи для первой точки сбыта (Рисунок 38).

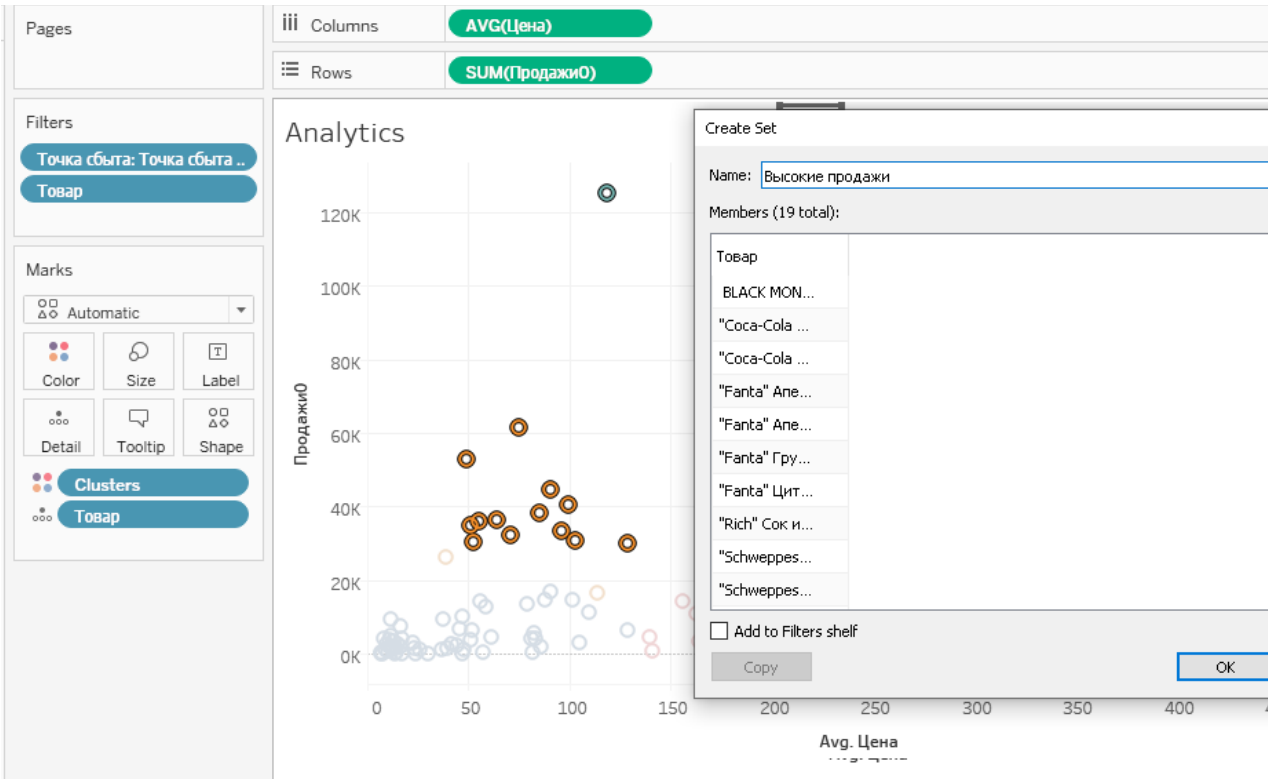

Рисунок 38 – создание набора данных высокие продажи

Добавим управление показом Высокие продажи на Marks, отобразим фильтрацию справа.

Покажем зависимость выручки от цены.

Для этого перенесём trend line на визуализированное поле (Рисунок 39).



Рисунок 39 – зависимость выручки от цены

Добавим средние оценки по осям с 95% отклонением. И переключим выручку на отображение average (Рисунок 40).

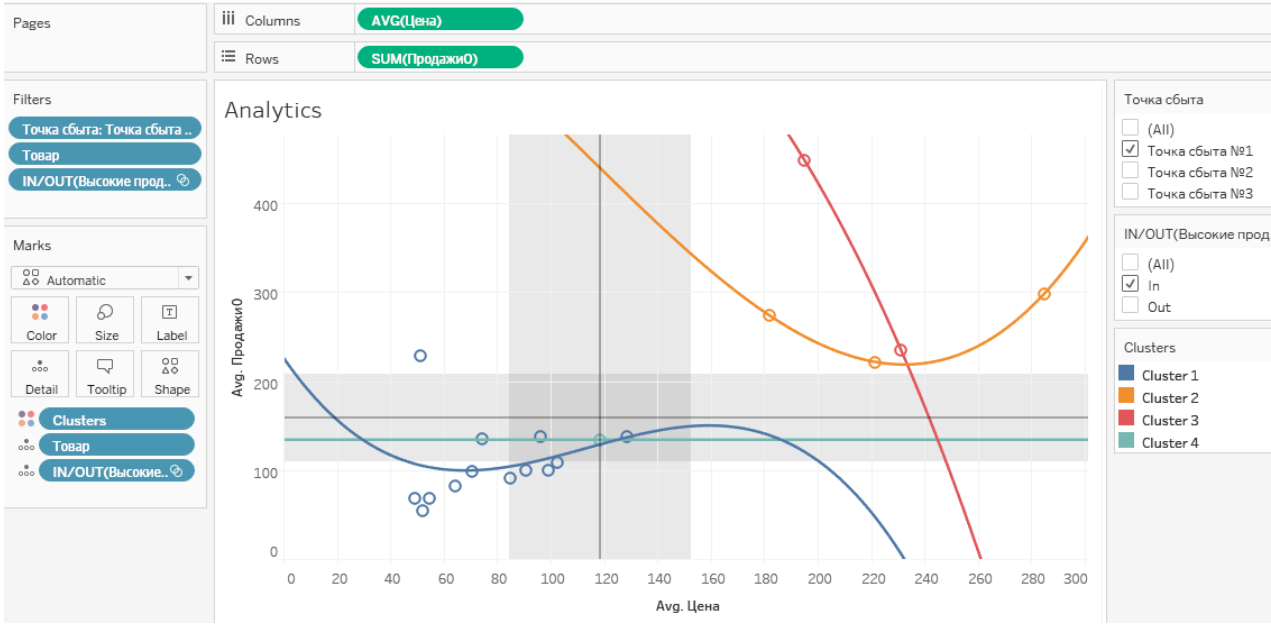

Рисунок 40 – определение средней цены и выручки

Средняя цена по первой точке сбыта по высокодоходным товарам с 95% интервалом составляет 118,5 рублей, выручка 158,9 рублей.

В результате интеллектуального анализа данных методом кластеризации были исследованы каждая из групп товаров с разделением на кластеры, благодаря чему были выявлены высокодоходные, средне доходные и низкодоходные товары. По каждому из кластеров была дана характеристика и рекомендация. Дополнительно был создан набор данных – высокодоходные товары по первой точке сбыта и отображена зависимость выручки от цены, а также определена средняя цена и выручка по новой группе данных.

Далее воспользуемся элементом аппарата технического анализа – trend line, для выявления тенденций изменения цен на группах товаров.

Выведем на визуализацию линии тренда для каждой исследуемой группы по каждому кластеру, чтобы понять какая тенденция формируется в пределах выбранной группы товаров. Для этого перенесём Trend Line в область построения визуализации и определим тип тренда «Power». Добавлении линии тренда изображено на Рисунок 41.

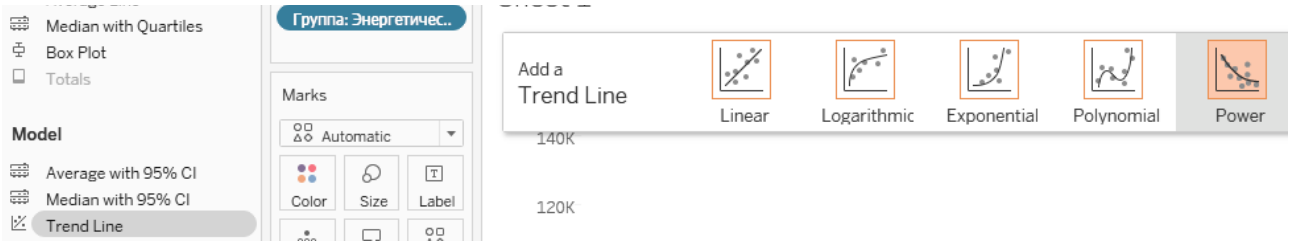

Рисунок 41 – добавление линии тренда

Получившийся результат отображён на Рисунок 42.

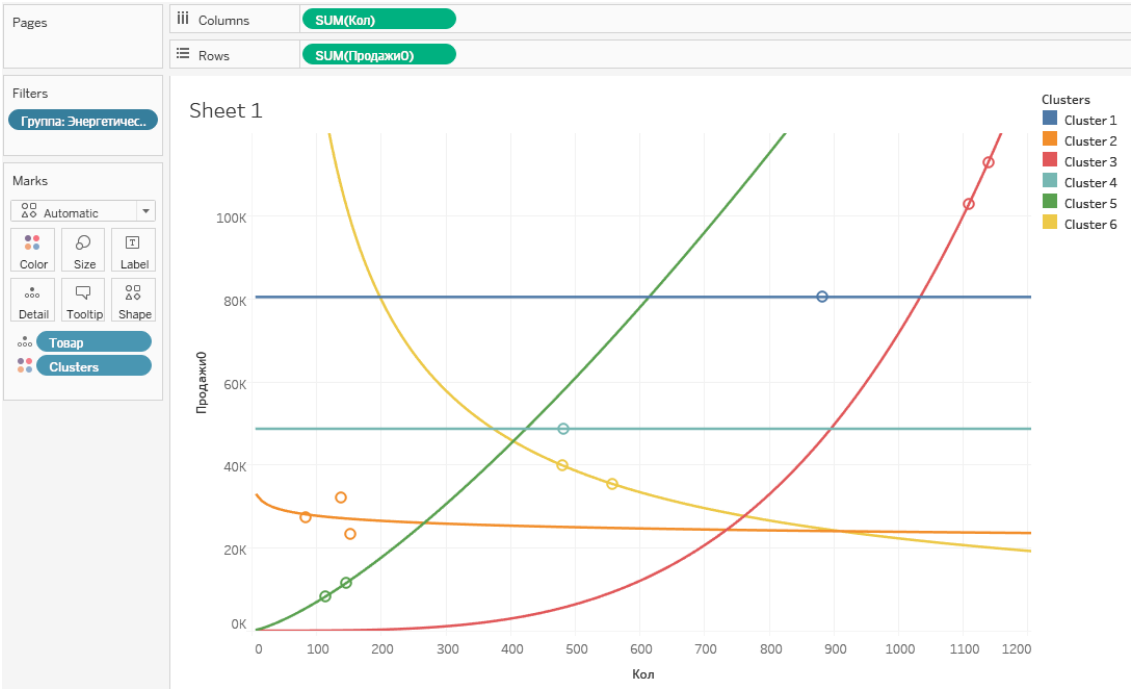

Рисунок 42 – линии тренда на точечной диаграмме

Для более полного понимания представленных данных выведем на диаграмму среднее медианное значение и 95% доверительный интервал для конкретного кластера, чтобы оценить область возможных изменений продаж. Для этого перенесем Median with 95% Cl на диаграмму. Анализ проведем по показателю Sum(Продажи). Иллюстрация действий представлена на Рисунок 43, Рисунок 44.

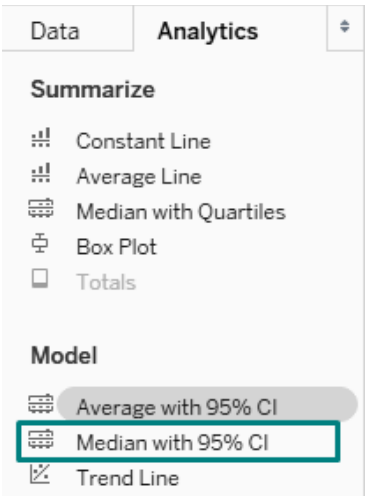

Рисунок 43 – добавление среднего медианного значение с 95% доверительным интервалом

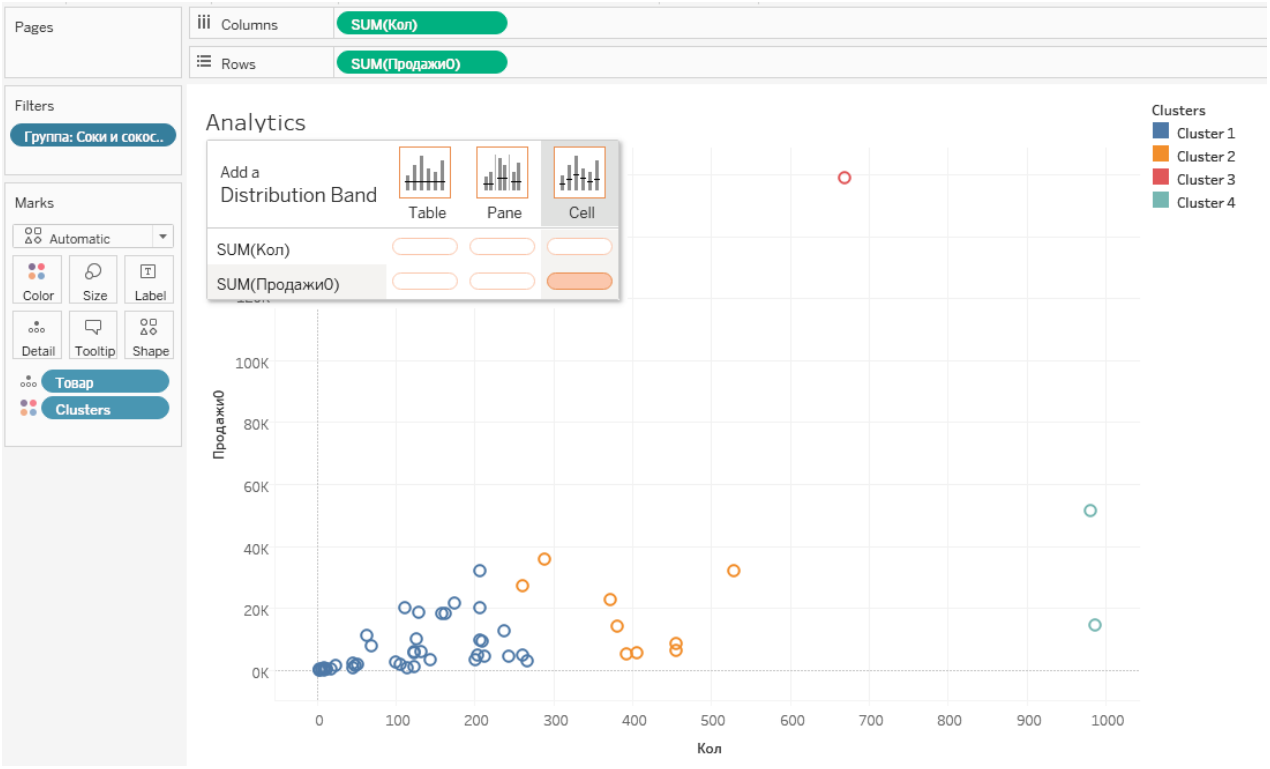

Рисунок 44 – добавление среднего медианного значение с 95% доверительным интервалом

Выбираем третий кластер. Медиана 158855 unit. Отображено на Рисунок 45.

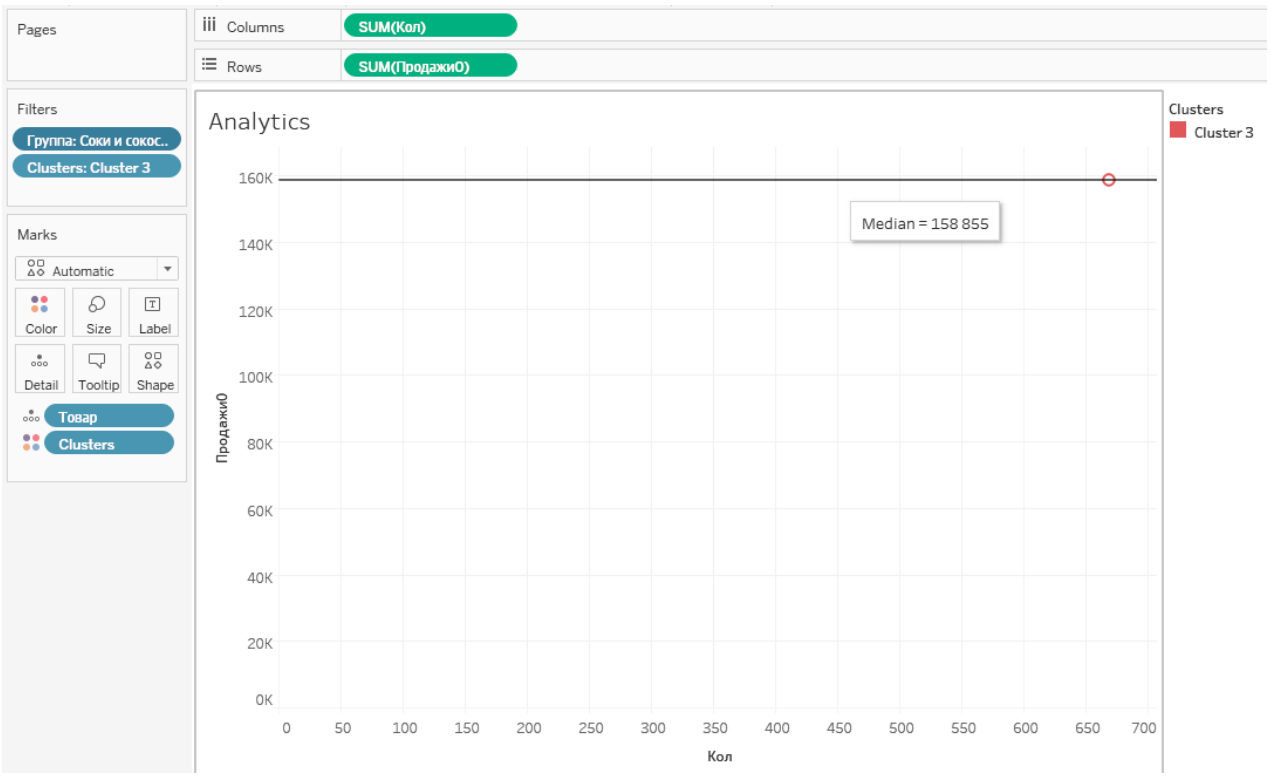

Рисунок 45 – отображение медианы третьего кластера

Переключимся на четвёртый кластер. Медиана – 32966 unit. Показано на Рисунок 46.



Рисунок 46 – отображение медианы первого кластера

Кластер номер 2. Медиана – unit. Проиллюстрировано на Рисунок 47.

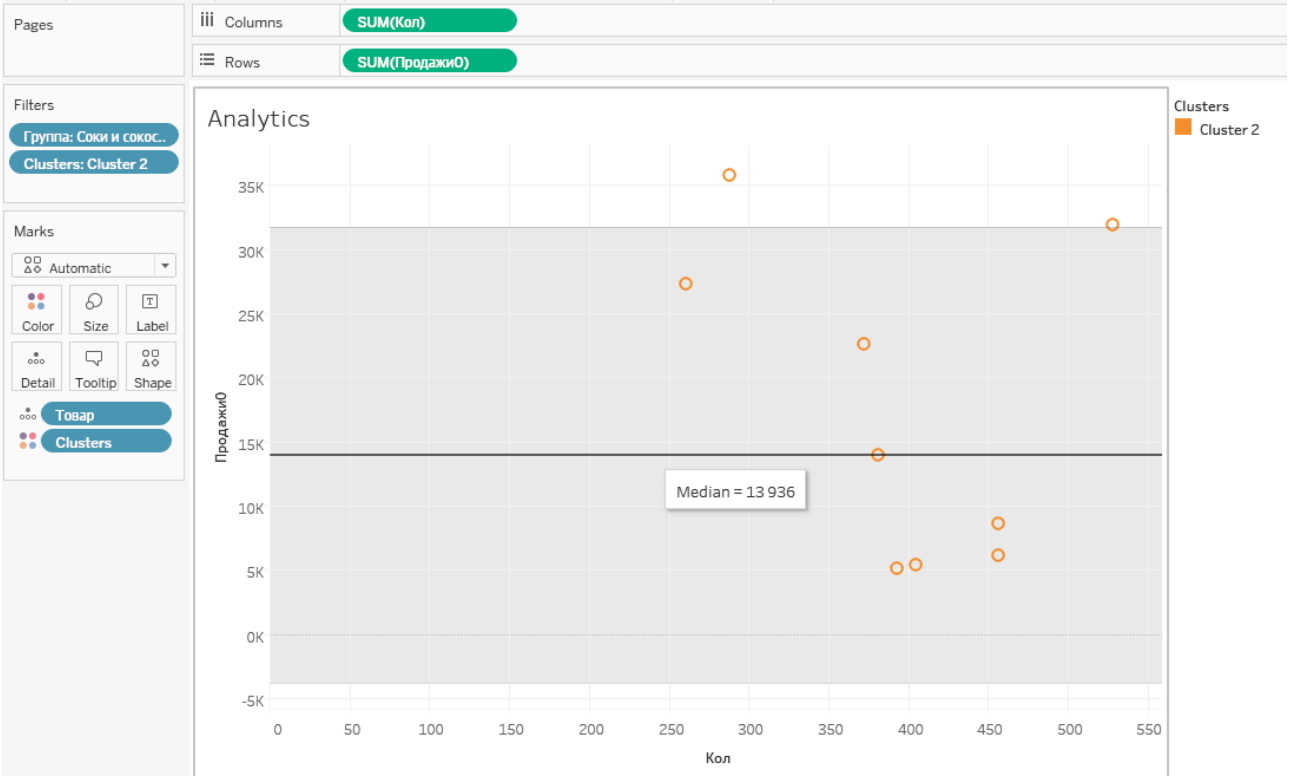

Рисунок 47 – отображение медианы второго кластера

Первый кластер. Медиана – 3389 unit. Выведем описание параметров линии трейда. Изображено на Рисунок 48.

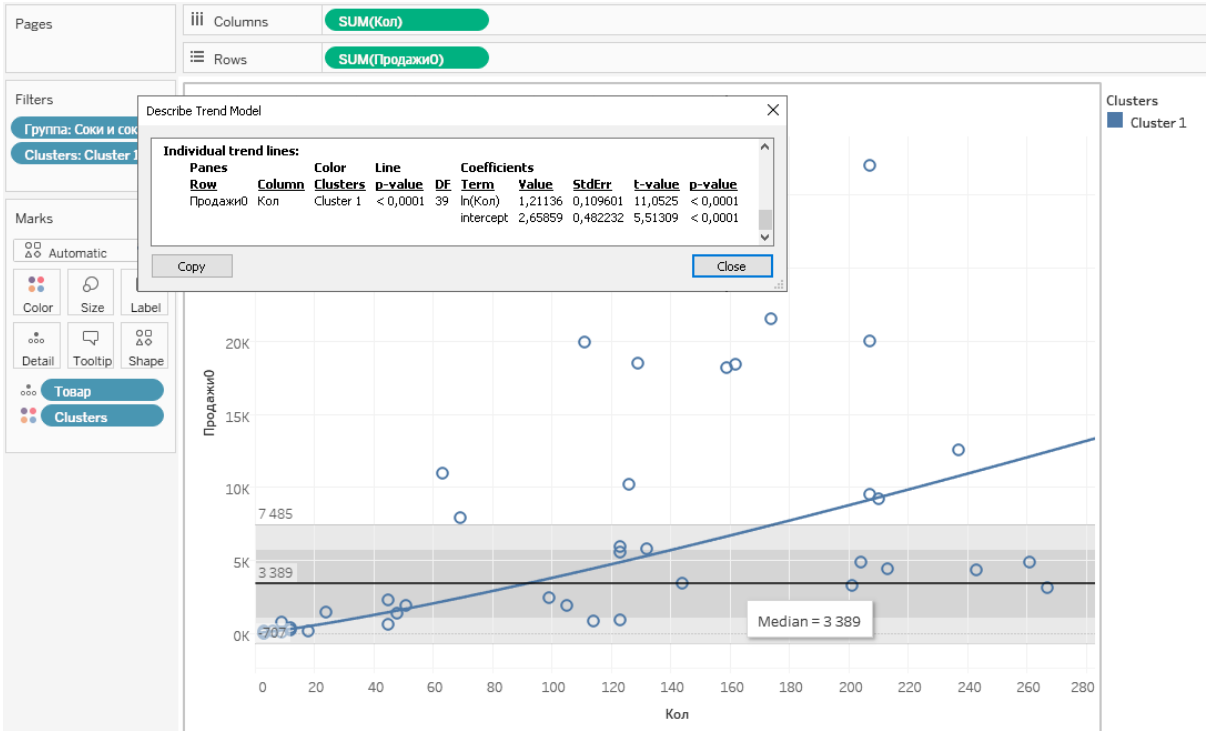

Рисунок 48 – отображение медианы и линии тренда с её описанием первого кластера

Используя инструмент trande line, а также Median with 95% Cl, были определены тенденции изменения цены и медианный доход по каждому кластеру товарной группы «Соки и сокосодержащие напитки».

Следующим этапом выполним анализ динамики продаж, исследуя который можно определить сильные и слабые места продаваемых товаров в зависимости от сезона.

3.3 АНАЛИЗ ДИНАМИКИ ПРОДАЖ

Выполним анализ динамики продаж по двум параметрам Sum(Продажи) и Sum(Кол) в разрезе «Месяцы» для товарной группы «Соки и сокосодержащие напитки». Для этого перенесём поле «Дата» на полку «Columns» и переключимся на «Month», поле Продажи и поле Кол на полку Rows. Поле Группа на полку Filters. «Measure Names» на полку «Marks». Вид графика – dual lines. Trade line – Power. Результат создания показан на Рисунок 50.

Создадим вторую ось для выражения Sum(Кол). Необходимый пункт для активации отображён на Рисунок 49.

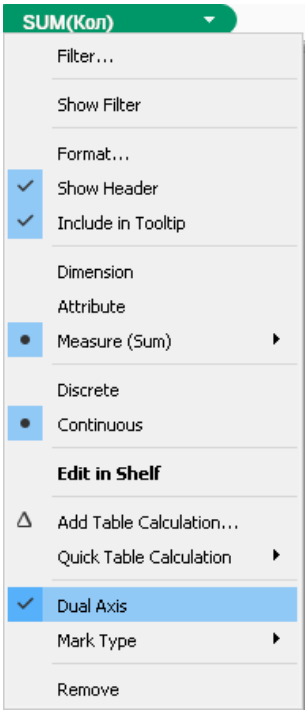

Рисунок 49 – создание двойной оси, пунктом "Dual Axis"

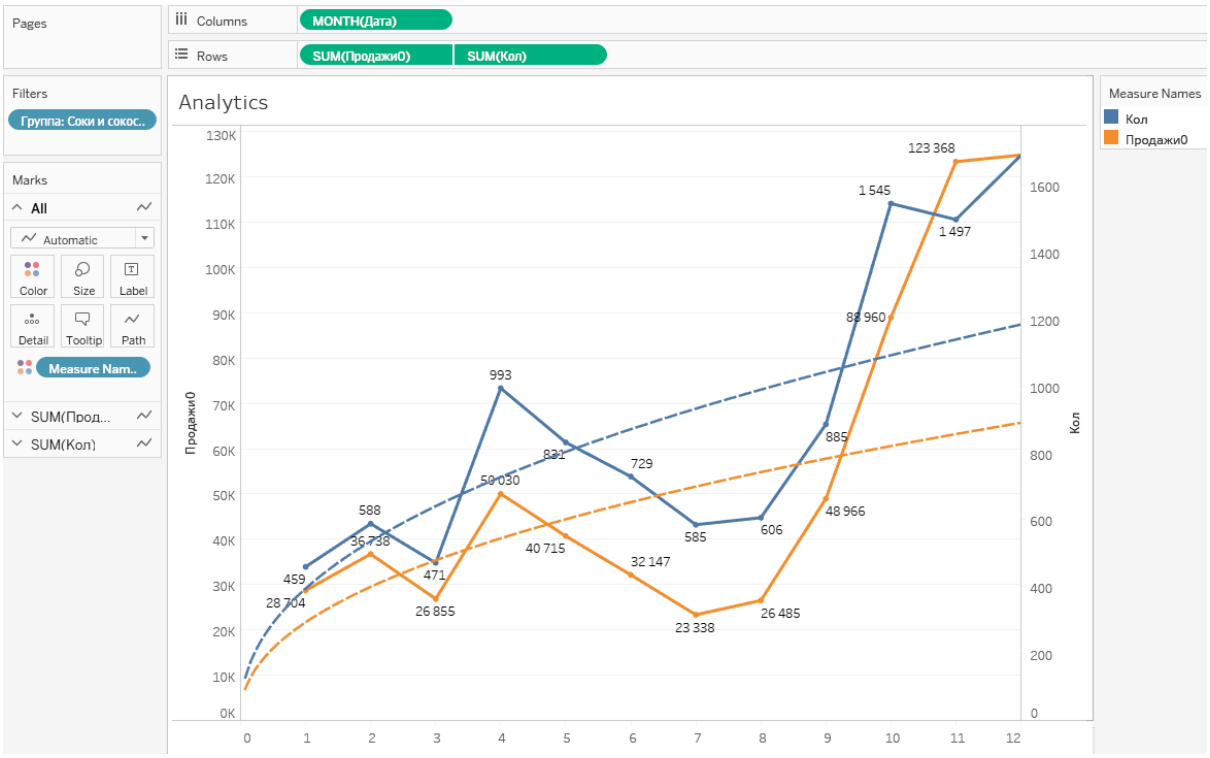

Рисунок 50 – анализ динамики продаж товарной группы "Соки и сокосодержащие напитки"

Согласно динамике, показанной на рисунке 50, наблюдаем спад продаж по обоим показателям в летний период. Продажи в межсезонье повышаются. Резкий подъём продаж в осенне-зимний период (до декабря включительно) нуждается в дополнительном исследовании. Это обусловливает общий повышающий тренд по сети.

Далее проведём сравнительный анализ продаж точки сбыта №1 с продажами по всей сети. Перенесём поле «Точки сбыта» на полку «Filters», отфильтруем ненужное, оставив первую точку. Двойной график изображён на Рисунок 51.

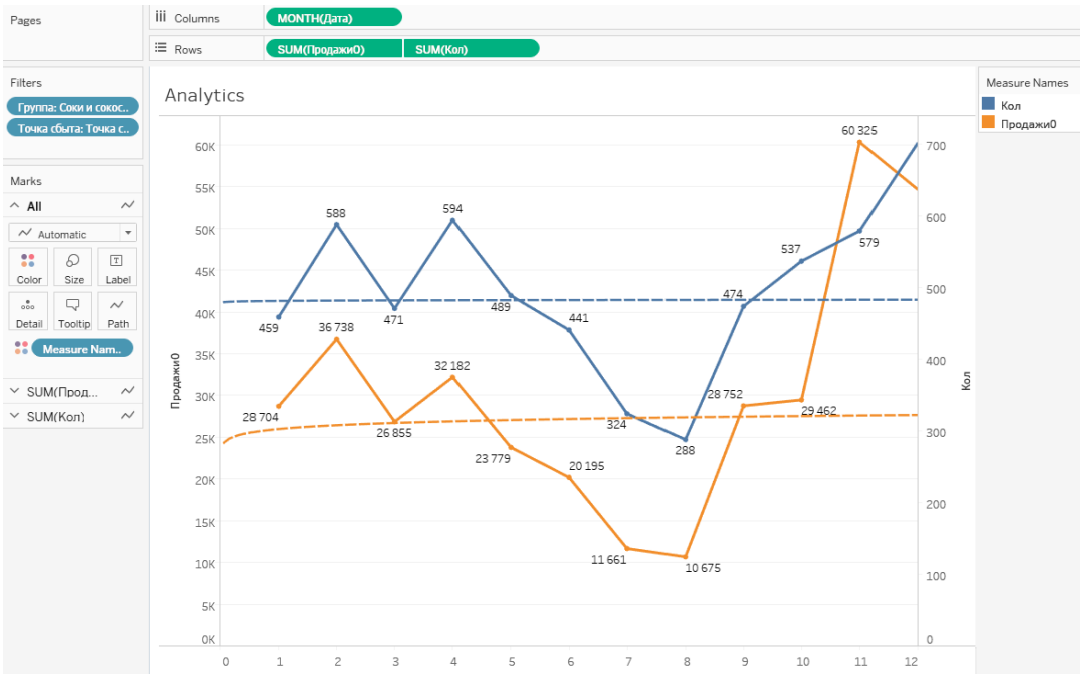

Рисунок 51 – анализ динамики продаж товарной группы "Соки и сокосодержащие напитки" на точке сбыта №1

Анализируя характер изменения и показатели продаж для данной группы товаров, можно предположить, что, точки сбыта имеют существенные различия в динамике продаж осенне-зимнего периода и вносят различный вклад в формирование общей динамики всей сети.

Выполним аналогичные действия по группе «Газированные напитки». График показан на Рисунок 52.

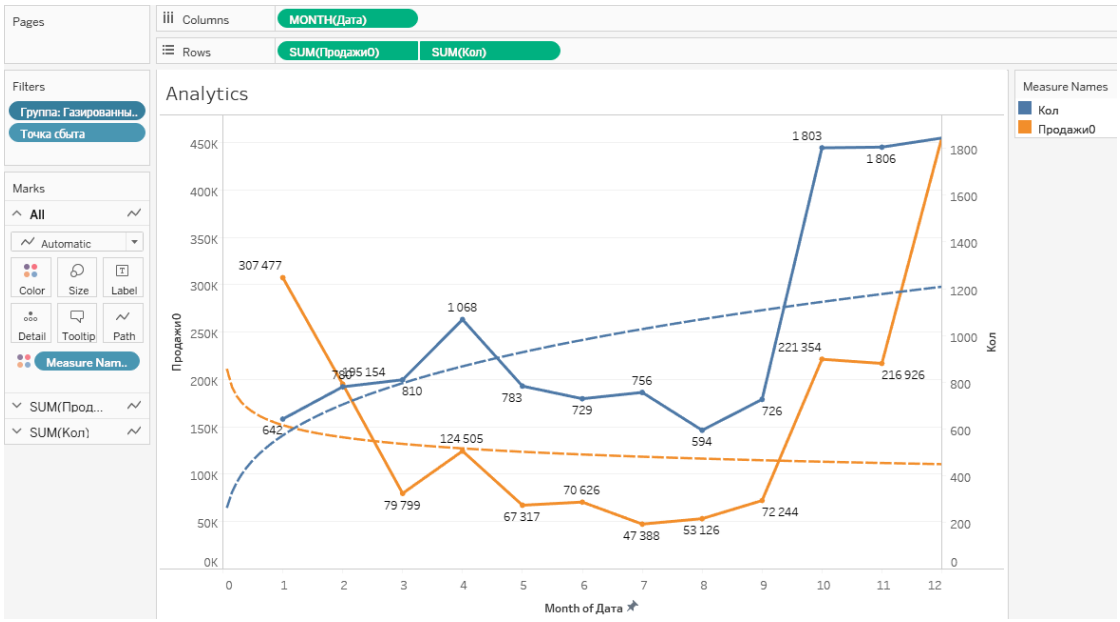

Рисунок 52 – анализ динамики продаж товарной группы "Газированные напитки"

Зимне-весенний период выражен спадом доходов, и незначительным ростом количества, что означает покупку дорогих товаров реже, а дешёвых в большем количестве. Весенне-летний показывает спад. Осенне-зимний выражен значительным подъемом, объем и выручка значительно выше, чем у рассмотренной выше торговой группы.

Изучая диаграмму, изображённую на Рисунок 53, наблюдаем изменение тренда с возрастающего для всей сети, к падающей для точки сбыта №1. В основном это обусловлено падением с января по март и низкими продажами в период с марта по ноябрь. Это необходимо учитывать при планировании закупок товара для данной точки сбыта.

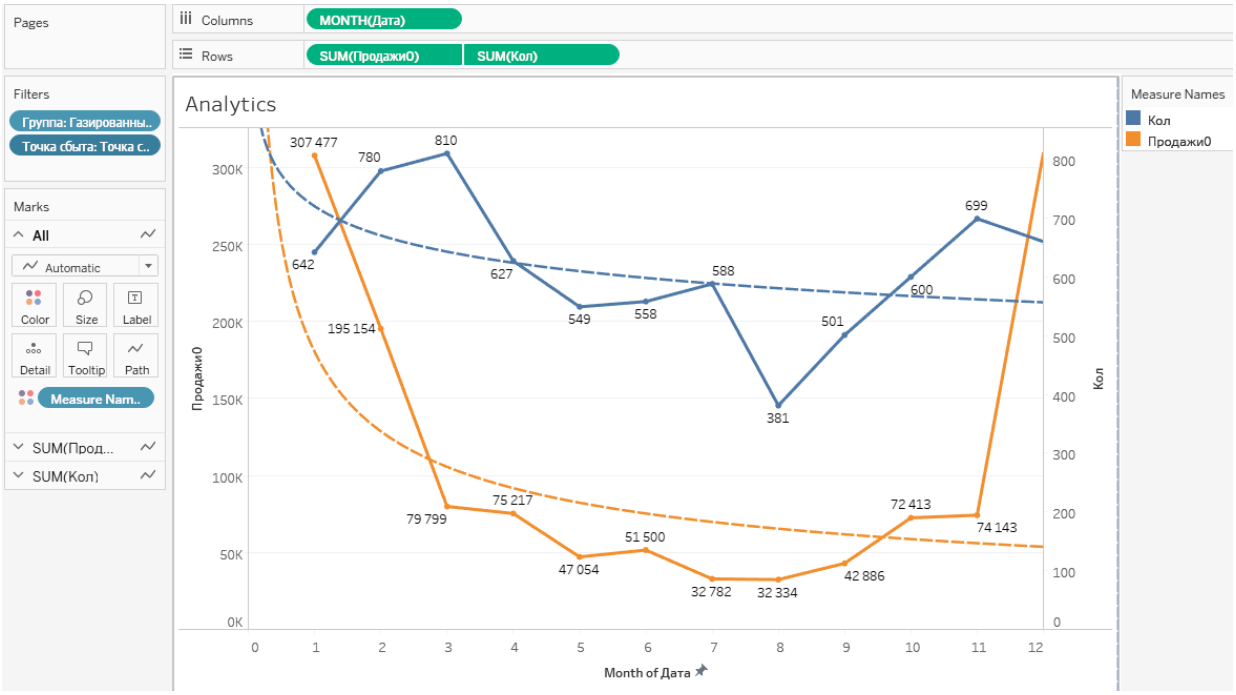

Рисунок 53 – анализ динамики продаж товарной группы "Газированные напитки" по точке сбыта №1

Следующая анализируемая группа по динамике продаж – «Энергетические напитки» отображена на Рисунок 54.

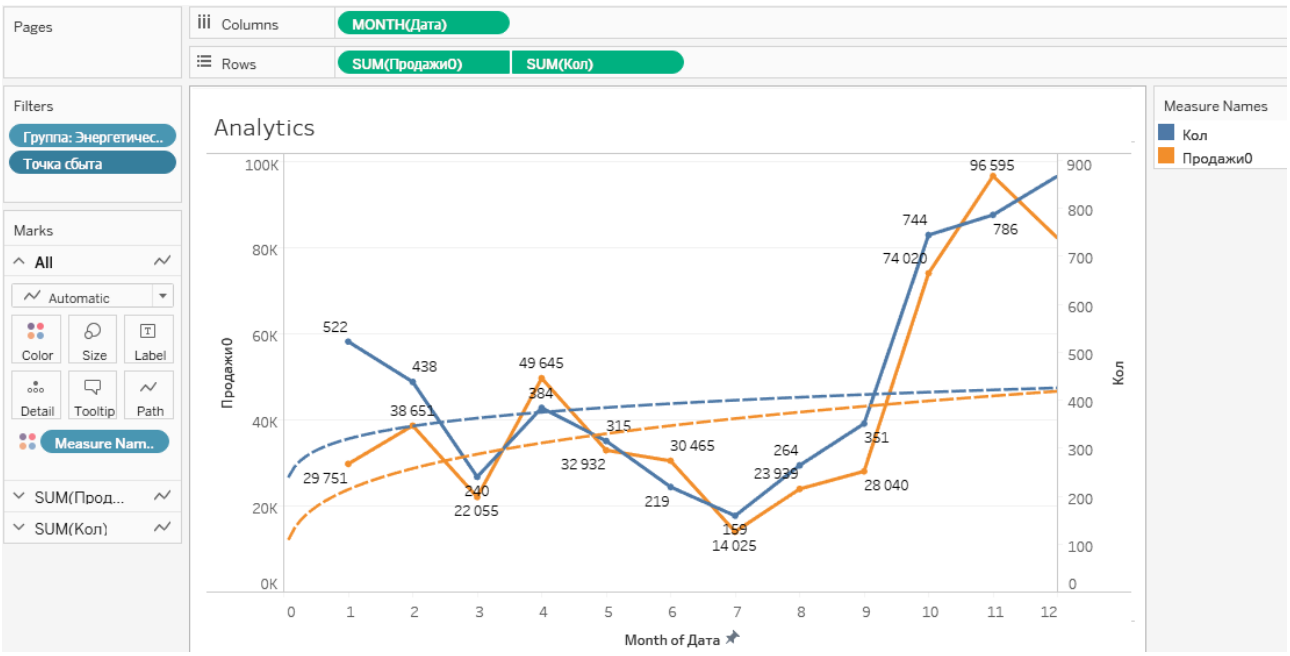

Рисунок 54 – анализ динамики продаж товарной группы "Энергетические напитки"

В целом, динамика продаж групп «Энергетические напитки» и «Сок и сокосодержащие напитки» похожи. Подход к управлению продаж этих двух групп можно скоординировать.

Динамика продаж для первой точки сбыта изображена на Рисунок 55.

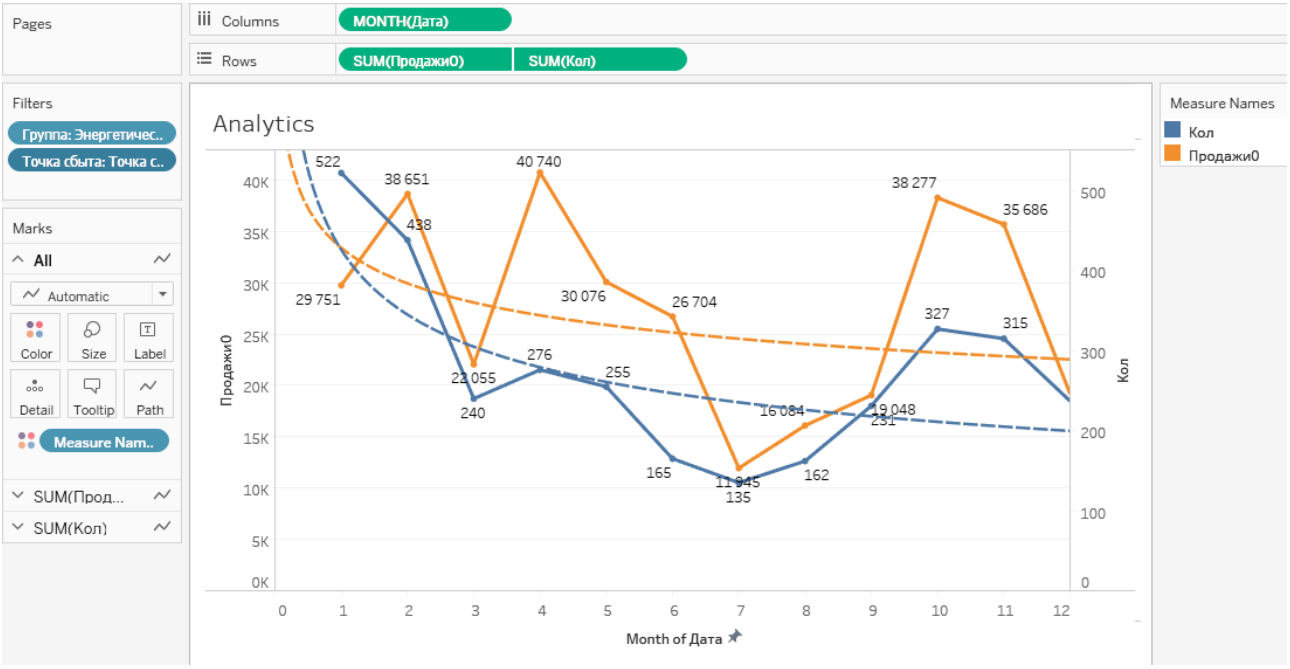

Рисунок 55 – анализ динамики продаж товарной группы "Энергетические напитки" по точке сбыта №1

В данном случае наблюдаем небольшое снижение в летний период. Начало года остаётся неизменным, что говорит о том, что данная точка является единственной, кто приносил доход. Такое возможно, когда две другие точки были открыты после этого времени. Доход в последние месяцы сокращается существенно.

Дополнительно рассмотрим часовую динамику. Выполним анализ продаж для точки сбыта №1 по трём группам товаров, временной интервал с 8 по 21 час. Графики для групп «Газированные напитки», «Соки и сокосодержащие напитки» и «Энергетические напитки» отображены на Рисунок 56,Рисунок 57,Рисунок 58 соответственно.

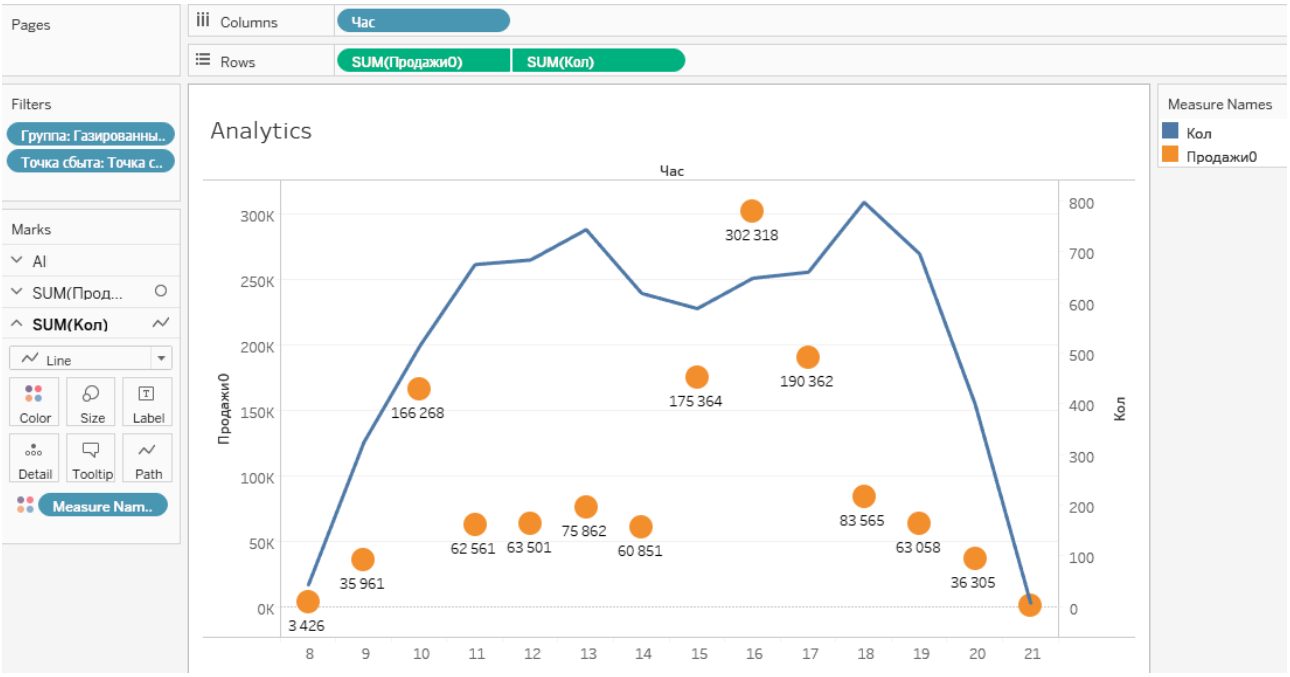

Рисунок 56 – анализ часовой динамики продаж товарной группы "Газированные напитки" для точки сбыта №1

Самые прибыльные часы продаж для данной группы товаров в точке сбыта №1, в порядке убывания: 16, 17, 15, 10. Низкодоходные часы в порядке возрастания: 8, 21, 9, 20. В остальное время выручка держится примерно на одном уровне.

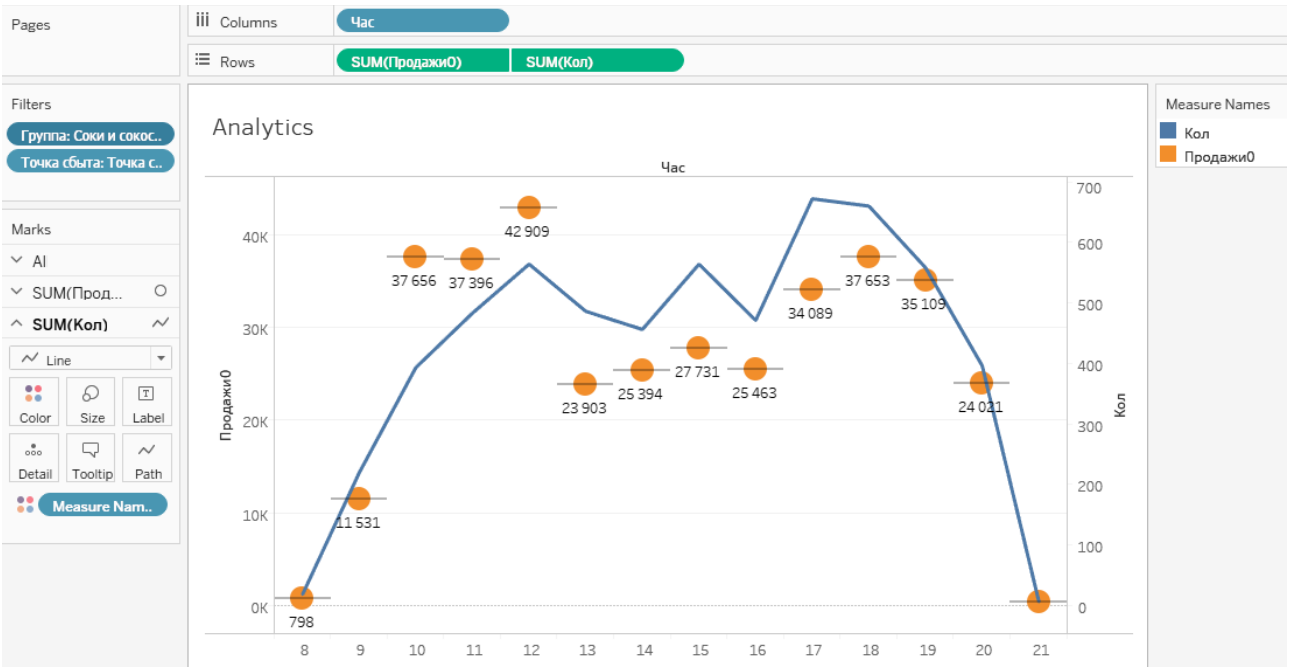

Рисунок 57 – анализ часовой динамики продаж товарной группы "Соки и сокосодержащие напитки" для точки сбыта №1

Для данной товарной группы самыми прибыльными часами в порядке убывания являются: 12, 11, 10, 18, 19, 17. Неприбыльные часы в порядке возрастания: 21, 8, 9. В остальное время доход держится на среднем уровне.

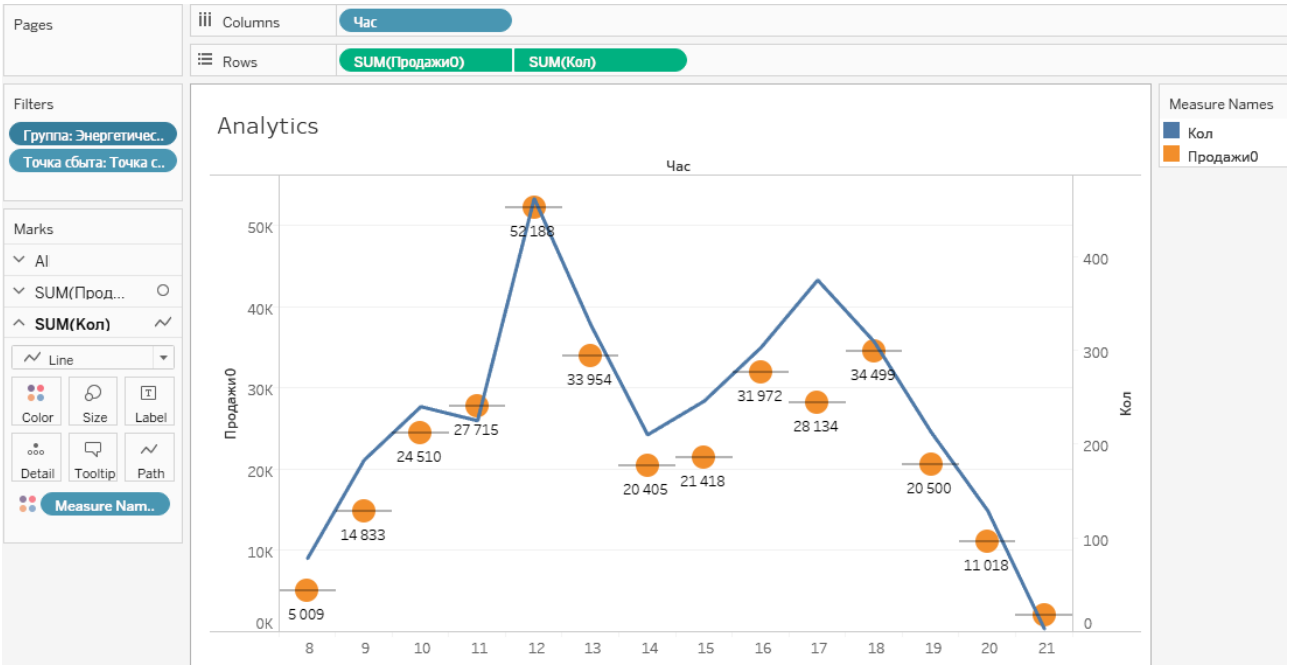

Рисунок 58 – анализ часовой динамики продаж товарной группы "Энергетические напитки" для точки сбыта №1

В данном случае имеем четыре малоприбыльных часа: 21, 8, 20, 9. И один высокодоходный – 12. Выше среднего: 18,13 и 16. Другое время в целом на среднем уровне.

Сводная диаграмма продаж в часовом разрезе по всем точкам сбыта для товарной группы «Энергетические напитки» отображена на Рисунок 59.

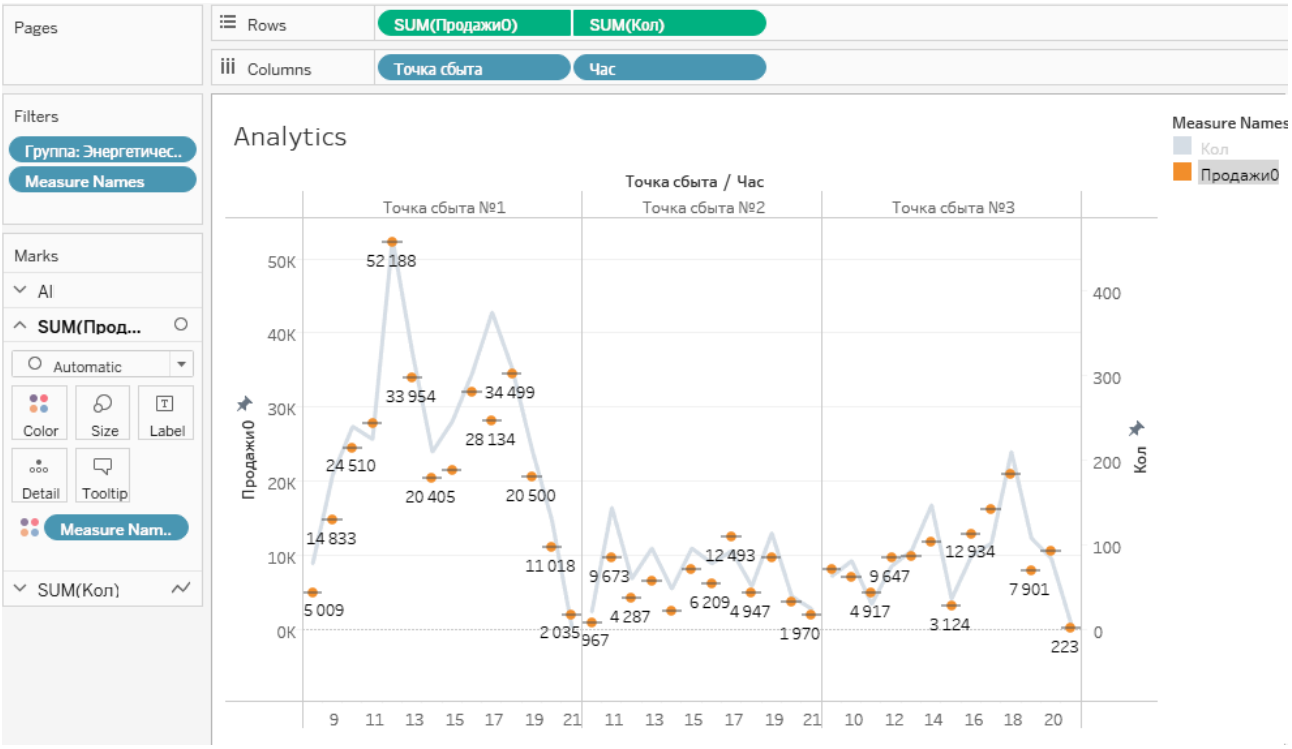

Рисунок 59 – сводная диаграмма продаж в часовом разрезе по всем точкам сбыта для товарной группы "Энергетические напитки"

В данном параграфе была исследована динамика продаж по каждой группе товаров на точке сбыта №1, определен спад и повышение продаж по сезонам, а также выявлены неприбыльные, прибыльные и такие часы, при котором доход умеренный. Проведено сравнение по часам продаж для одной из групп среди всех точек сбыта.

3.4 ПРОГНОЗИРОВАНИЕ

Создадим прогноз по товару «Black Monster Assault» из группы высокодоходных товаров в первой точке сбыта. Для этого на полку Columns перенесём: группу, товар, дату (установим по месяцам). На полку Rows добавим

выручку. На полку Filters добавим код группы, точку сбыта с сортировкой по первой и высокодоходные товары (Рисунок 60).

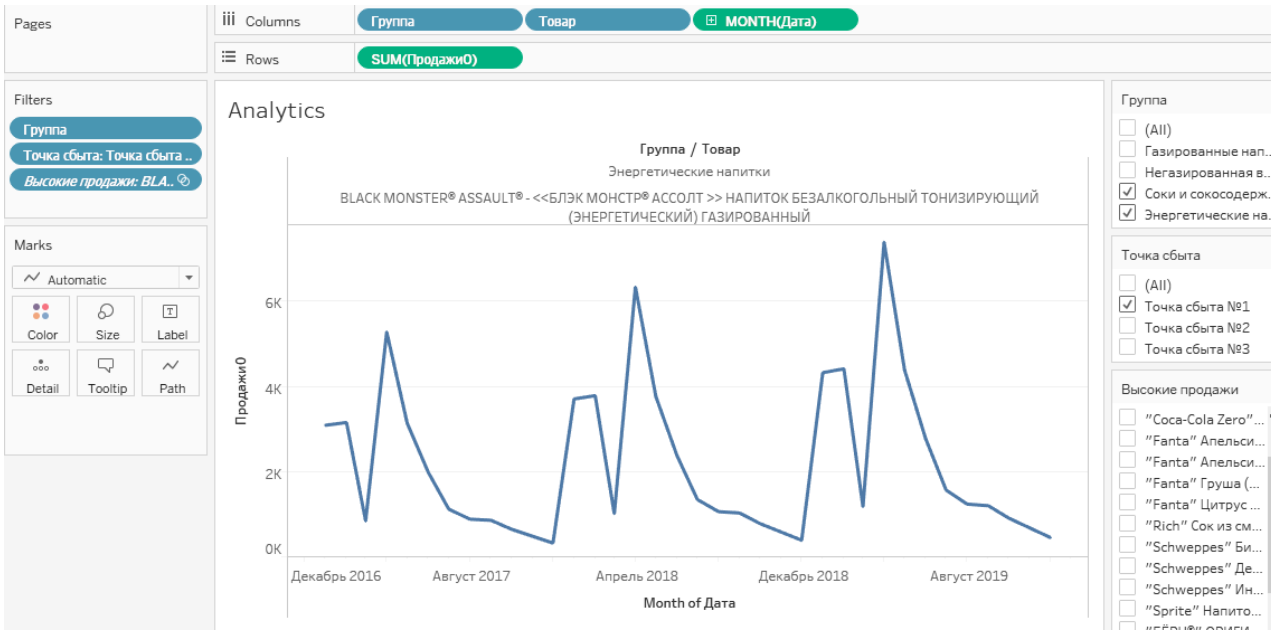

Рисунок 60 – помесячная выручка выбранного товара в течение трёх лет

Добавляем forecast в область viz и включим Mark Lavel по всем точкам (Рисунок 61).

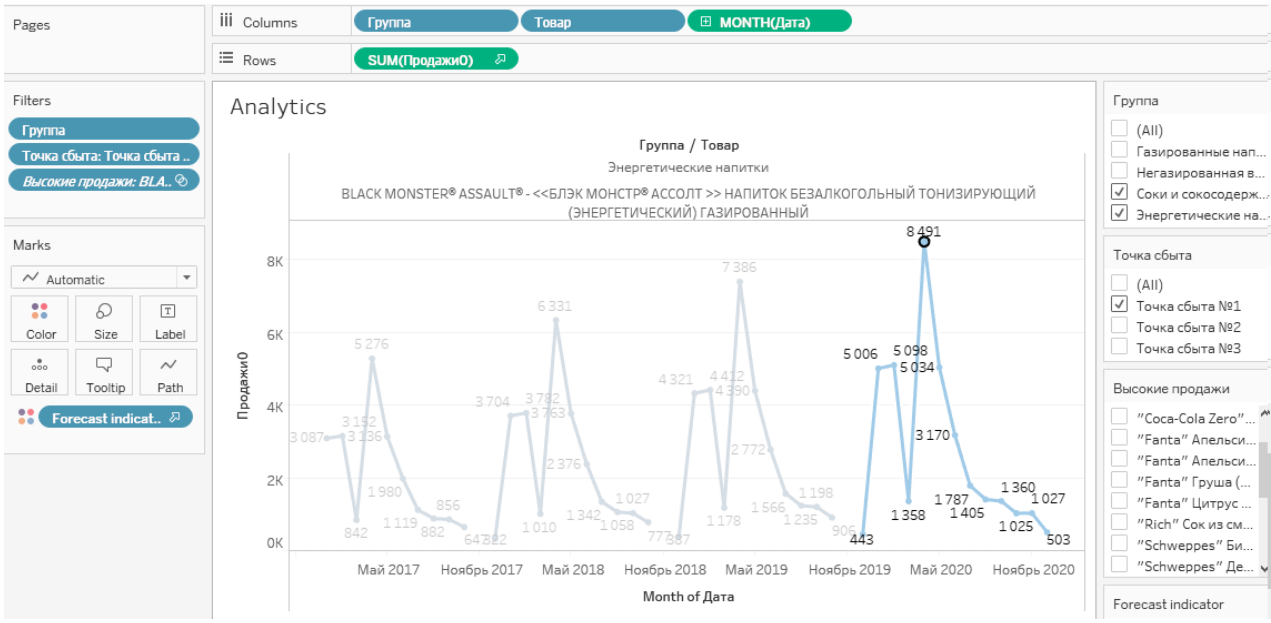

Рисунок 61 – помесячный прогноз выручки на следующий год

Проведём аналогичные действия для товара Powerade Снежное золото (Рисунок 62).

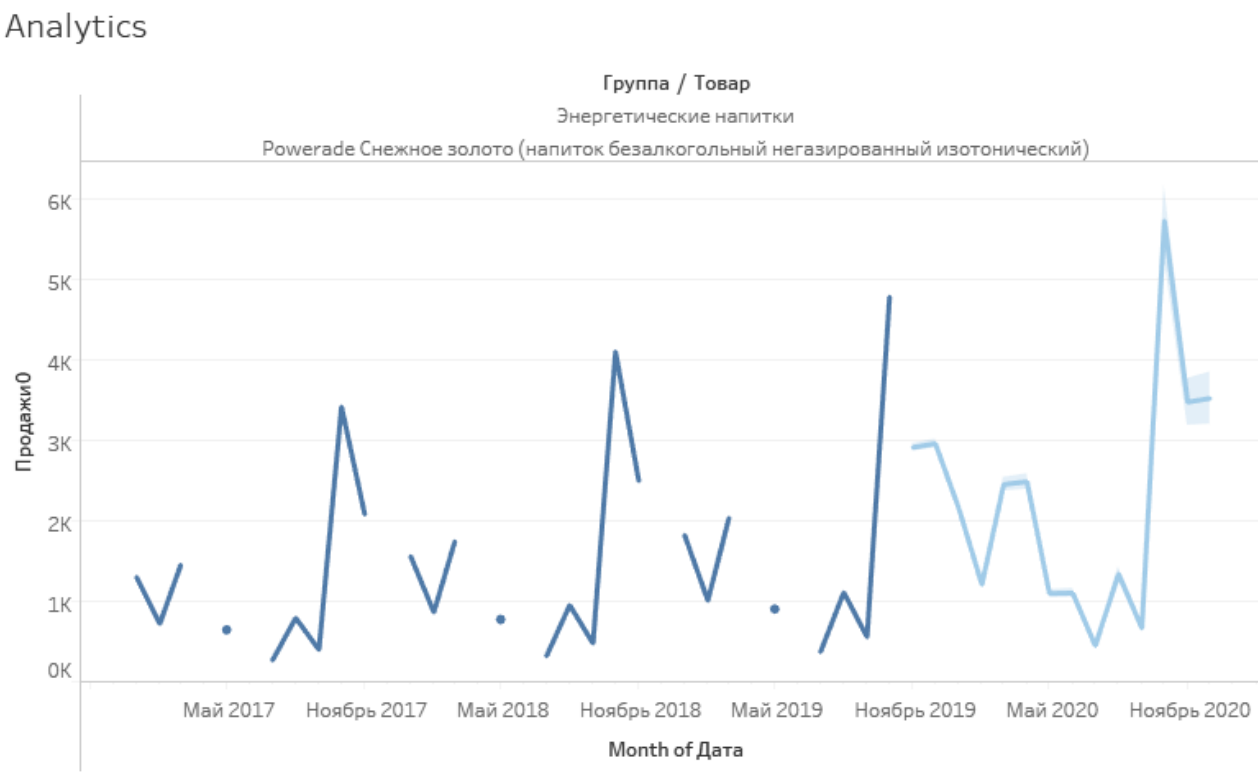

Рисунок 62 – помесячный прогноз выручки выбранного товара на следующий год

Здесь мы можем наблюдать тенденцию прекращения продаж товара на протяжении трёх лет в одно и то же время, а именно декабрь, апрелю и июнь. При анализе продаж ранее, было выдвинуто предположении о том, что точка сбыта только была открыта, однако это не так. Для уточнения ситуации лучше всего обратиться к специалистам компании.

Построим прогноз для данного товара на следующие два года. Для этого откроем Forecast Options, в Forecast Length установим значение 3 в поле Until (Рисунок 63).

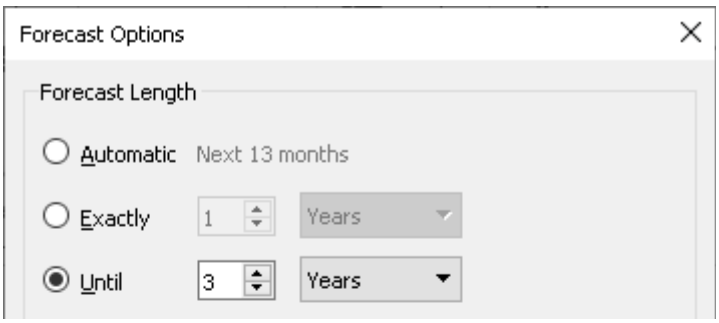

Рисунок 63 – настройка forecast

В результате получаем прогноз на два года. Ситуацию отображает Рисунок 64.

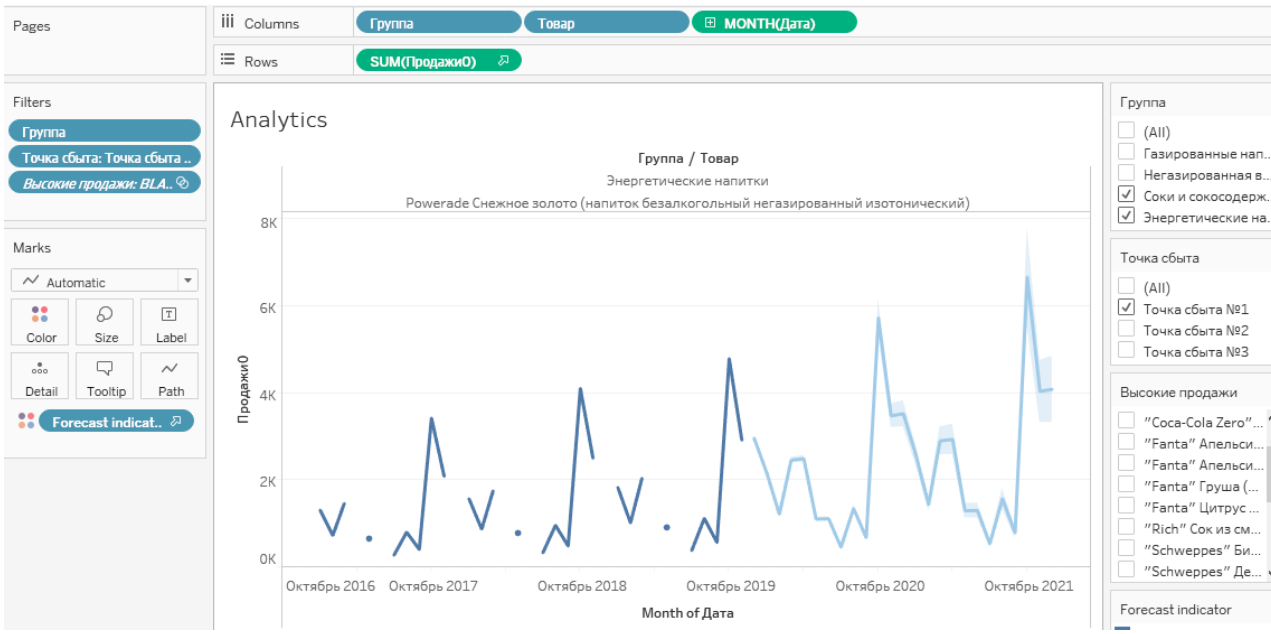

Рисунок 64 – прогноз на 2 года

Теперь мы можем наблюдать прогноз товара «Powerade Снежное золото» на три года без сбоев, с отображением допустимого отклонения от прогноза. При устранении сбоев продаж ожидается увеличение выручки примерно на 7103 денежных единиц на следующий год и на 8298д.е. на 2021 год соответственно.

В данном параграфе был рассмотрен прогноз по двум товарам от одного до двух лет в разрезе месяц, определили у одного из товаров сбой в продажах и подсчитали вероятную выручку при устранении данной проблемы.

3.5 КОМПЛЕКСНЫЙ АНАЛИЗ ПРОДАЖ

Приступим к созданию dashboard. Создаём новый лист dashboard (Рисунок 65).

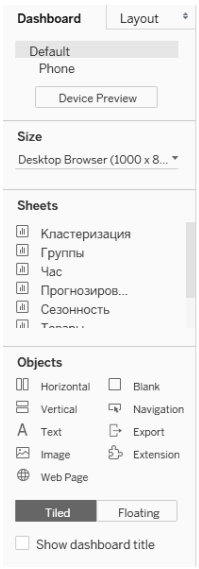

Рисунок 65 – новый лист dashboard

Добавим sheets: товары, сезонность, час, группы, точки сбыта. Увеличим размер отображения до (1700x900), удобно расположим sheets и свяжем их фильтром. Отображено на Рисунок 66.

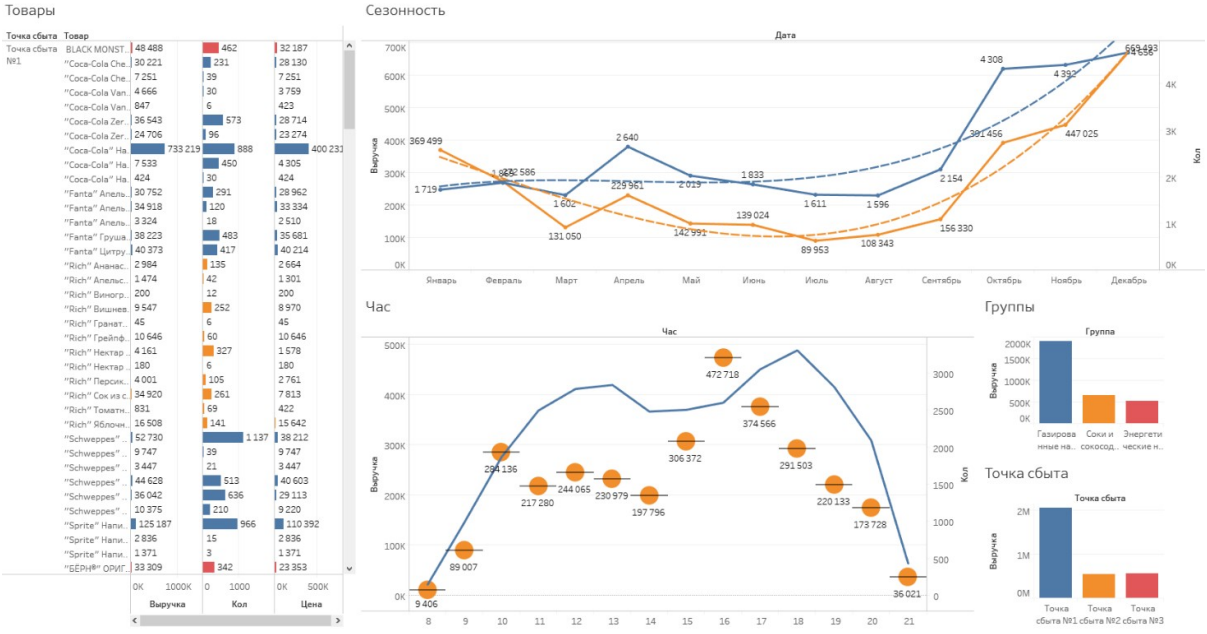

Рисунок 66 – dashboard с данными

Dashboard состоит из sheets, поэтому для того, чтобы он получился удобный и красивый, необходимо также настроить каждый sheets.

С помощью связанных по фильтру sheets, dashboard позволяет выставлять точечную настройку, что в свою очередь, помогает глубже изучить и понять данные, а затем провести комплексный анализ.

Проведём анализ. Нас интересует летний период товарной группы «Энергетические напитки» (Рисунок 67).

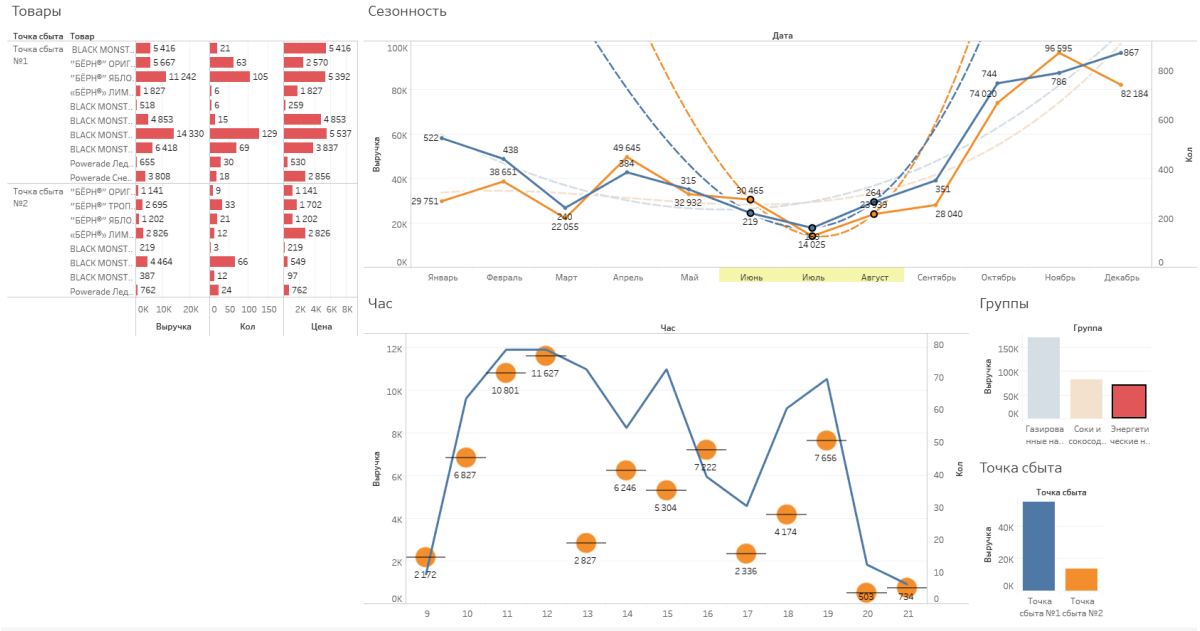

Рисунок 67 – анализ энергетических напитков в летний период

Из всех периодов, летний для данной товарной группы считается самым низкодоходным, выручка по которому составила 68429 д.е., из них на первую точку сбыта приходится 54733 д.е., на вторую 13696 д.е. Самые прибыльные часы в порядке убывания: 12, 11, 19, 16; низкодоходные часы в порядке возрастания: 20, 21, 9, 17, 13. Средне доходные по убыванию: 10, 14, 15, 18.

Самые прибыльные товары: Black Monster Assault и «Бёрн» яблоко-киви, выручка которых составила 14330 д.е. и 11242 д.е. соответственно. Низкодоходные товары в первой точке сбыта: Black Monster The doctor TM, Powerade Ледяная буря, выручка 518 д.е. и 655 д.е. соответственно. Во второй точке сбыта: Black Monster, Black Monster Khaos, Powerade Ледяная буря, выручка 219 д.е., 387 д.е. и 762 д.е. соответственно.

Разберём продажи одного из вышеперечисленных низкодоходных товаров, например Powerade Ледяная буря (Рисунок 68).

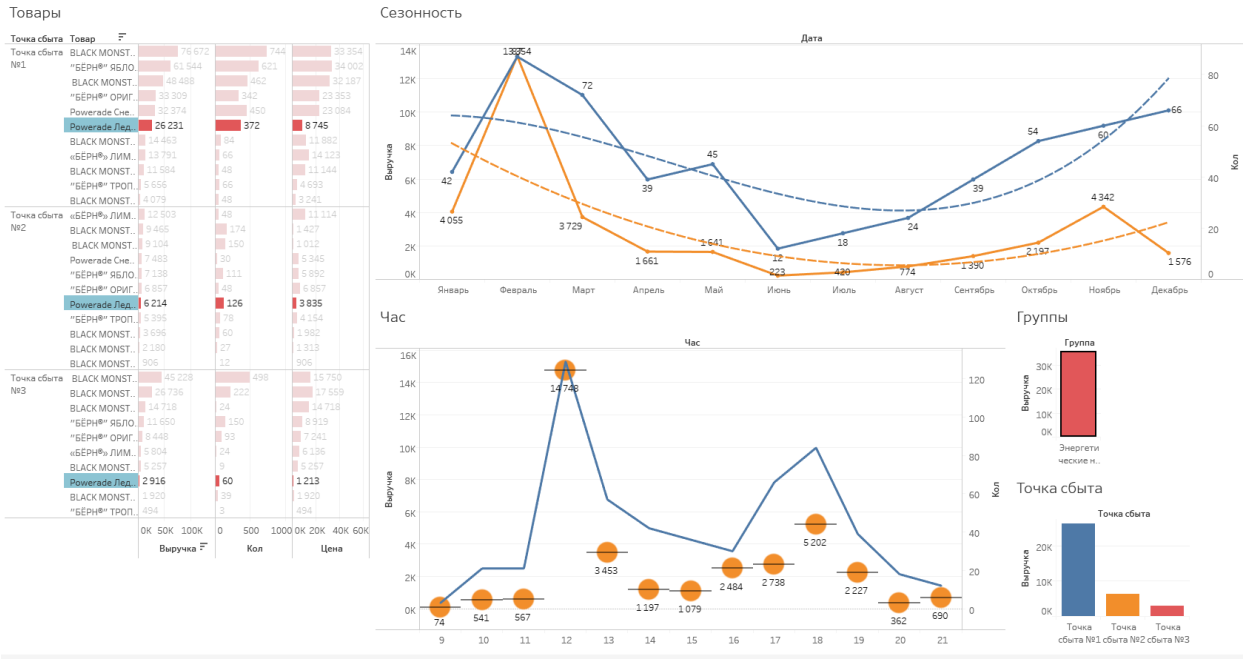

Рисунок 68 – анализ продаж Powerade Ледяная буря

В весенний и летний период выручка идёт на спад, в осенний растёт, в зимний также растёт. За всё время выручка составила 35362 д.е., из которых на первую, вторую и третью приходятся 26231 д.е., 6214 д.е. и 2,916 д.е. соответственно. Самый популярный час покупки – 12, на втором месте –18, в остальные покупали умеренно, кроме 9 часа – он является самым низкодоходным.

Для того, чтобы увеличить популярность и выручку данного товара, необходимо повысить его известность, путём рекламы, а также чаще делать всевозможные скидки, акции, например два по цене одного.

В данном параграфе был создан dashboard, проанализирована одна товарная группа по летнему периоду, выявлены низкодоходные товары, один из которых был взят для анализа. Были выявлены: сезонность товара, выручка по каждой из точек сбыта; популярное, умеренное и непопулярное время покупки товара в разрезе часа. Дана рекомендация для увеличения продаж данного товара.

ЗАКЛЮЧЕНИЕ

В рамках выполнения выпускной квалификационной работы были рассмотрены технологии интеллектуального анализа, изучены и выбрано инструментальное средство Tableau, специализирующееся на визуализации данных. Подробно рассмотрены используемые методы кластеризации и визуализации данных.

В результате интеллектуального анализа данных методом кластеризации были исследованы каждая из групп товаров с разделением на кластеры, благодаря чему были выявлены высокодоходные, средне доходные и низкодоходные товары. По каждому из кластеров была дана характеристика и рекомендация. Дополнительно был создан набор данных – высокодоходные товары по первой точке сбыта и отображена зависимость выручки от цены, а также определена средняя цена и выручка по новой группе данных. Используя инструмент trande line, а также Median with 95% Cl, были определены тенденции изменения цены и медианный доход по каждому кластеру одной из товарных групп.

Проведено сравнение по часам продаж для одной из групп среди всех точек сбыта. Рассмотрен прогноз по двум товарам от одного до двух лет в разрезе месяц, определён сбой в продажах у одного из товаров и подсчитана вероятная выручка при устранении данной проблемы.

Создали dashboard, проанализировав одну товарную группу по летнему периоду, выявлены низкодоходные товары, один из которых был взят для анализа. Были выявлены: сезонность товара, выручка по каждой из точек сбыта; популярное, умеренное и непопулярное время покупки товара в разрезе часа. Дана рекомендация для увеличения продаж данного товара.

Объектом исследования являлись методы и инструментальные средства интеллектуального анализа данных. Предмет исследования – методы кластеризации и визуализации в анализе продаж.

Таким образом результатом выполнения выпускной квалификационной работы стала разработка рекомендаций на основе данных, предоставленных организацией. Проблемы, обозначенные и рассмотренные в выпускной квалификационной работе, решены. Это позволяет сделать вывод, что задание выполнено полностью.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. А.А. Шумейко, С.Л. Сотник Интеллектуальный анализ данных Введение в Data Mining. – Днепропетровск: Белая Е.А., 2012. – 212 с
2. Барсегян А. А., Куприянов М. С, Степаненко В. В., Холод И. И., Технологии анализа данных. Data Mining, Visual Mining, Text Mining, OLAP [Текст]: учеб. - Изд. 2-е издание: БХВ-Петербург, 2007.-79 с.
3. Дюк В.А., Самойленко А.П. Data Mining [Текст]: учебный курс. - СПб.: Питер, 2010.-С.67-87.
4. Орешков В.И., Паклин Н.Б. Бизнес-аналитика. От данных к знаниям [Текст]: - учеб. - Спб.: Питер, 2013. С.105-110.
5. Чубукова И.А. Курс лекций по Data Mining [Текст]: – учеб. - М.: Интернет-университет информационных технологий, Бином. Лаборатория знаний, 2010. С.152-179.
6. Кокорева, Я. В. Поэтапный процесс кластерного анализа данных на основе алгоритма кластеризации k-means / Я. В. Кокорева, А. А. Макаров. — [Текст]: непосредственный // Молодой ученый. — 2015. — № 13 (93). — С. 126-128.
7. Левкович-Маслюк. Л. Интервью Григория Пятецкий−Шапиро журналу "Компьютерра" [Электронный ресурс] / Л.Левкович. Маслюк. — Электрон. журн. — Компьютерра, 2007. — Режим доступа к журн.: https://www.kdnuggets.com/gpspubs/piatetsky-interview-computerra.pdf. — Электрон. версия печ. публикации, 11 (679), стр. 50
8. Введение в добычу данных (Data Mining) [Электронный ресурс] / Богатырёв С., Симонова А. — Электрон. текстовые дан. — 2006. — Режим доступа: https://logic.pdmi.ras.ru/~yura/internet/01ia-seminar-note.pdf, свободный. — Электрон. версия печ. публикации
9. Визуализация данных // Microsoft Power BI [Электронный ресурс] /. — Электрон. текстовые дан. — Режим доступа:

https://powerbi.microsoft.com/ru-ru/, свободный (дата обращения: 22.05.2020).

10. Деревья решений - общие принципы работы // BaseGroup Labs [Электронный ресурс] /. — Электрон. текстовые дан. — Режим доступа: https://basegroup.ru/community/articles/description, свободный (дата обращения: 18.05.2020).
11. Логистическая регрессия и ROC-анализ // BaseGroup Labs [Электронный ресурс] /. — Электрон. текстовые дан. — Режим доступа: https://basegroup.ru/community/articles/logistic, свободный (дата обращения: 18.05.2020).
12. Национальный открытый университет «Интуит» [Электронный ресурс]. – Режим доступа: http://www.intuit.ru/studies/courses/6/6/lecture/200?page=1, свободный, (дата обращения: 20.05.2020).
13. Нейросеть // BaseGroup Labs [Электронный ресурс] /. — Электрон. текстовые дан. — Режим доступа: https://basegroup.ru/deductor/function/algorithm/neuronet, свободный (дата обращения: 18.05.2020).
14. Никульская, Н.А. Задачи, модели и методы Data Mining [Электронный ресурс] / Н.А. Никульская. — Электрон. текстовые дан. — Режим доступа: http://elib.sfu-kras.ru/bitstream/handle/2311/20642/Nikul'skaya_(2).pdf. — Электрон. версия статьи
15. Обзор методов Data Mining [Электронный ресурс] /. — Электрон. текстовые дан. — 2017. — Режим доступа: https://ai-news.ru/2017/08/obzor_metodov_data_mining_intellektualnyj_analiz_dannyh.html
16. Портал для студентов «Студми» [Электронный ресурс]. – Режим доступа: http://studme.org/1813022725076/menedzhment/sistemy_upravleniya_bazami_dannyh_naborom_algoritmov_data_mining, свободный, (дата обращения: 20.05.2020).

17. Data Mining — добыча данных // BaseGroup Labs [Электронный ресурс] /. —— Режим доступа: https://basegroup.ru/community/articles/data-mining, свободный (дата обращения: 16.05.2020).
18. Data Mining. Курс лекций // НОУ ИНТУИТ [Электронный ресурс] /. — Электрон. текстовые дан. — Режим доступа: https://www.intuit.ru/studies/courses/6/6/lecture/158, свободный (дата обращения: 16.05.2020).
19. Microsoft Power BI — набор средств бизнес-аналитики для анализа данных и предоставления ценной информации // BI CONSULT [Электронный ресурс] — Режим доступа: https://biconsult.ru/products/microsoft-power-bi-nabor-sredstv-biznes-analitiki-dlya-analiza-dannyh-i-predostavleniya, свободный (дата обращения: 26.05.2020).
20. QlikView — Powerful Interactive Analitics & Dashboards // Qlik [Электронный ресурс] — Режим доступа: https://www.qlik.com/us/products/qlikview, свободный (дата обращения: 22.05.2020).
21. QlikView и плагины - система бизнес анализа от экспертов // BI CONSULT [Электронный ресурс] — Режим доступа: https://biconsult.ru/products/qlikview, свободный (дата обращения: 22.05.2020).
22. Tableau - система интерактивной бизнес аналитики // BI CONSULT [Электронный ресурс] — Режим доступа: https://biconsult.ru/products/tableau, свободный (дата обращения: 23.05.2020).
23. Tableau Desktop // Tableau [Электронный ресурс] — Режим доступа: https://www.tableau.com/products/desktop, свободный (дата обращения: 23.05.2020).