Правительство Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»
(НИУ ВШЭ)

Московский институт электроники и математики им. А.Н. Тихонова

ПРАКТИЧЕСКАЯ РАБОТА № 4
по дисциплине «Информатика»

ТЕМА РАБОТЫ

«Анализ данных с использованием ассоциативных правил в RM»

Москва, 2024

Оглавление

1. Введение 2
2. Содержание практической работы 4
3. Ход работы 6
4. Приобретаемые навыки 24
5. Обобщенная задача для индивидуального варианта 24
6. Распределение вариантов 25

1. Введение

Целью данной лабораторной работы является освоение методов анализа данных и визуализации на основе ассоциативных правил с использованием инструмента RapidMiner. Ассоциативные правила позволяют выявлять взаимосвязи между объектами в больших наборах данных, что актуально для анализа потребительского поведения, оптимизации процессов и принятия стратегических решений.

В рамках работы студенты выполнят предобработку данных, выявят частые наборы элементов с использованием алгоритма FP-Growth, сформируют ассоциативные правила, а также проанализируют и визуализируют результаты. Особое внимание уделяется этапу предобработки данных, поскольку корректная обработка данных значительно влияет на точность и информативность полученных результатов.

Кроме того, студенты изучат подходы к анализу статистических характеристик данных с использованием блоков статистики.

2. Содержание практической работы

Описание работы:

В основе работы лежит набор данных о транзакциях в интернет-магазине, содержащий информацию о покупках клиентов. Данные включают множество атрибутов, таких как номера транзакций, названия товаров, их количество, цену и страну, в которой была совершена покупка.

Этапы выполнения работы:

1. Провести предобработку данных, включая фильтрацию транзакций, удаление ненужных столбцов и подготовку данных для анализа.
2. Сформировать ассоциативные правила на основе алгоритма FP-Growth.
3. Провести анализ полученных ассоциативных правил для выявления закономерностей в покупках клиентов.
4. Построить визуализации, отражающие частоту покупок и взаимосвязи между товарами.
5. Использовать статистический анализ для оценки характеристик набора данных.

О наборе данных:

Анализ проводится на наборе данных, содержащем следующие характеристики:

- **InvoiceNo** — уникальный идентификатор транзакции.
- **Description** — название товара.
- **Quantity** — количество купленных единиц товара.
- **UnitPrice** — стоимость одной единицы товара.
- **CustomerID** — идентификатор клиента.
- **Country** — страна, в которой была совершена покупка.

Ключевые особенности данных:

- **Количество записей:** более 540 тысяч.
- **Формат данных:** табличный набор, содержащий числовые и текстовые атрибуты.

- **Потенциальные проблемы:** наличие пропущенных данных, отрицательные значения в столбце Quantity, наличие транзакций, не относящихся к покупкам (возвраты).

Специфика работы:

Для выполнения лабораторной работы будут использоваться только следующие столбцы:

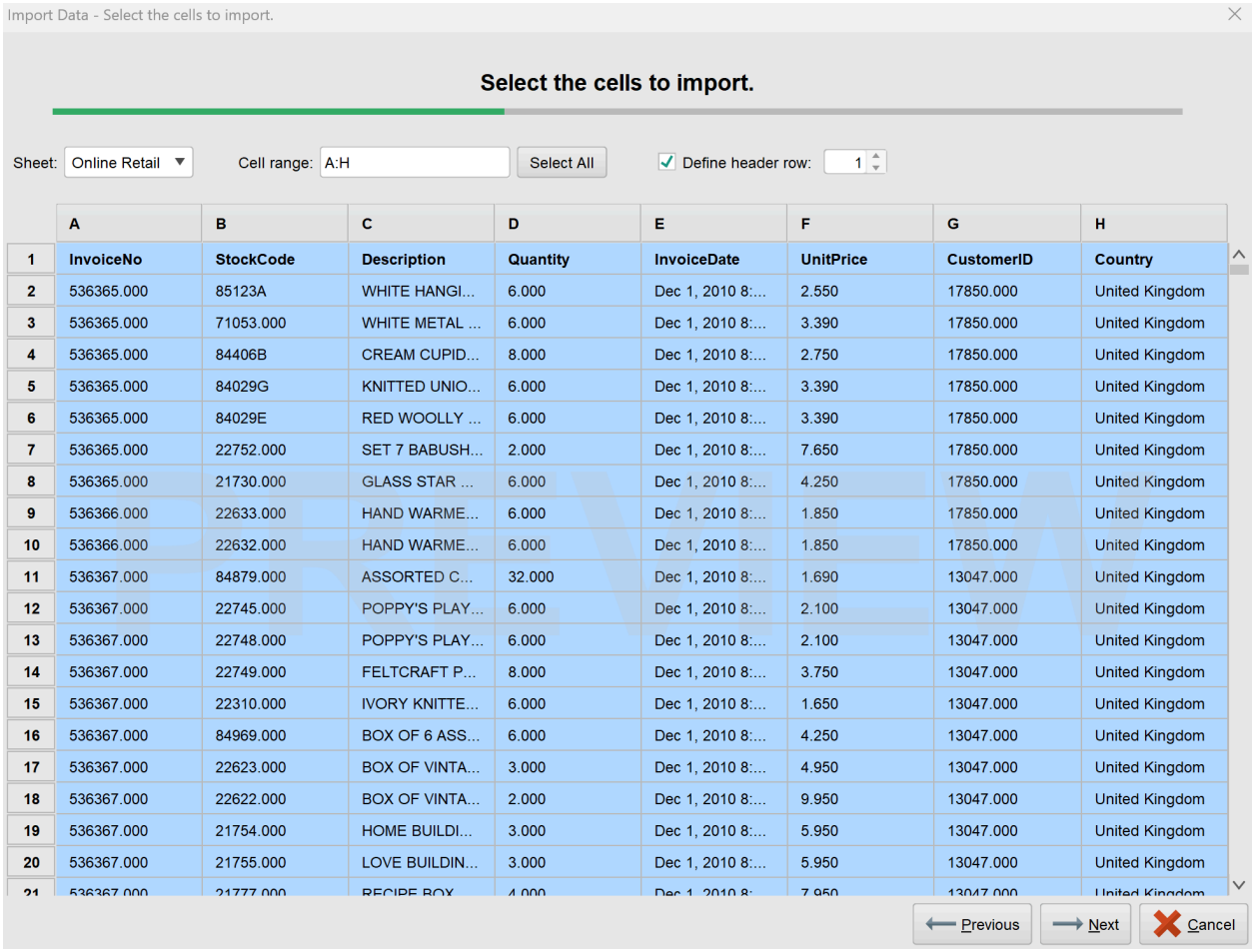
- **InvoiceNo** (идентификатор транзакции),
- **Description** (название товара).

3. Ход работы

Загрузка набора данных

1. Откройте RapidMiner Studio.
2. В главном меню выберите **"Create New Process"**.
3. Воспользуйтесь функцией **"Import data".**
4. Загрузите набор данных о транзакциях, выбрав файл **"Online Retail"** в формате xlsx.
5. Сохраните полученную базу данных в папку со своей работой.
6. Не забудьте поставить галочку у «replace errors with missing values»
7. В результате вы увидите таблицу с данными, содержащими атрибуты: (**InvoiceNo; Description; Quantity; UnitPrice; CustomerID; Country**)

Данные успешно загружены, их структура показана на рисунке 3.2.



	A	B	C	D	E	F	G	H
1	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
2	536365.000	85123A	WHITE HANGI...	6.000	Dec 1, 2010 8:...	2.550	17850.000	United Kingdom
3	536365.000	71053.000	WHITE METAL ...	6.000	Dec 1, 2010 8:...	3.390	17850.000	United Kingdom
4	536365.000	84406B	CREAM CUPID...	8.000	Dec 1, 2010 8:...	2.750	17850.000	United Kingdom
5	536365.000	84029G	KNITTED UNIO...	6.000	Dec 1, 2010 8:...	3.390	17850.000	United Kingdom
6	536365.000	84029E	RED WOOLLY ...	6.000	Dec 1, 2010 8:...	3.390	17850.000	United Kingdom
7	536365.000	22752.000	SET 7 BABUSH...	2.000	Dec 1, 2010 8:...	7.650	17850.000	United Kingdom
8	536365.000	21730.000	GLASS STAR ...	6.000	Dec 1, 2010 8:...	4.250	17850.000	United Kingdom
9	536366.000	22633.000	HAND WARME...	6.000	Dec 1, 2010 8:...	1.850	17850.000	United Kingdom
10	536366.000	22632.000	HAND WARME...	6.000	Dec 1, 2010 8:...	1.850	17850.000	United Kingdom
11	536367.000	84879.000	ASSORTED C...	32.000	Dec 1, 2010 8:...	1.690	13047.000	United Kingdom
12	536367.000	22745.000	POPPY'S PLAY...	6.000	Dec 1, 2010 8:...	2.100	13047.000	United Kingdom
13	536367.000	22748.000	POPPY'S PLAY...	6.000	Dec 1, 2010 8:...	2.100	13047.000	United Kingdom
14	536367.000	22749.000	FELTCRAFT P...	8.000	Dec 1, 2010 8:...	3.750	13047.000	United Kingdom
15	536367.000	22310.000	IVORY KNITTE...	6.000	Dec 1, 2010 8:...	1.650	13047.000	United Kingdom
16	536367.000	84969.000	BOX OF 6 ASS...	6.000	Dec 1, 2010 8:...	4.250	13047.000	United Kingdom
17	536367.000	22623.000	BOX OF VINTA...	3.000	Dec 1, 2010 8:...	4.950	13047.000	United Kingdom
18	536367.000	22622.000	BOX OF VINTA...	2.000	Dec 1, 2010 8:...	9.950	13047.000	United Kingdom
19	536367.000	21754.000	HOME BUILDI...	3.000	Dec 1, 2010 8:...	5.950	13047.000	United Kingdom
20	536367.000	21755.000	LOVE BUILDIN...	3.000	Dec 1, 2010 8:...	5.950	13047.000	United Kingdom
21	536367.000	21777.000	RECIPE BOX ...	4.000	Dec 1, 2010 8:...	7.950	13047.000	United Kingdom

Рисунок 3.1 – подготовка данных к выгрузке

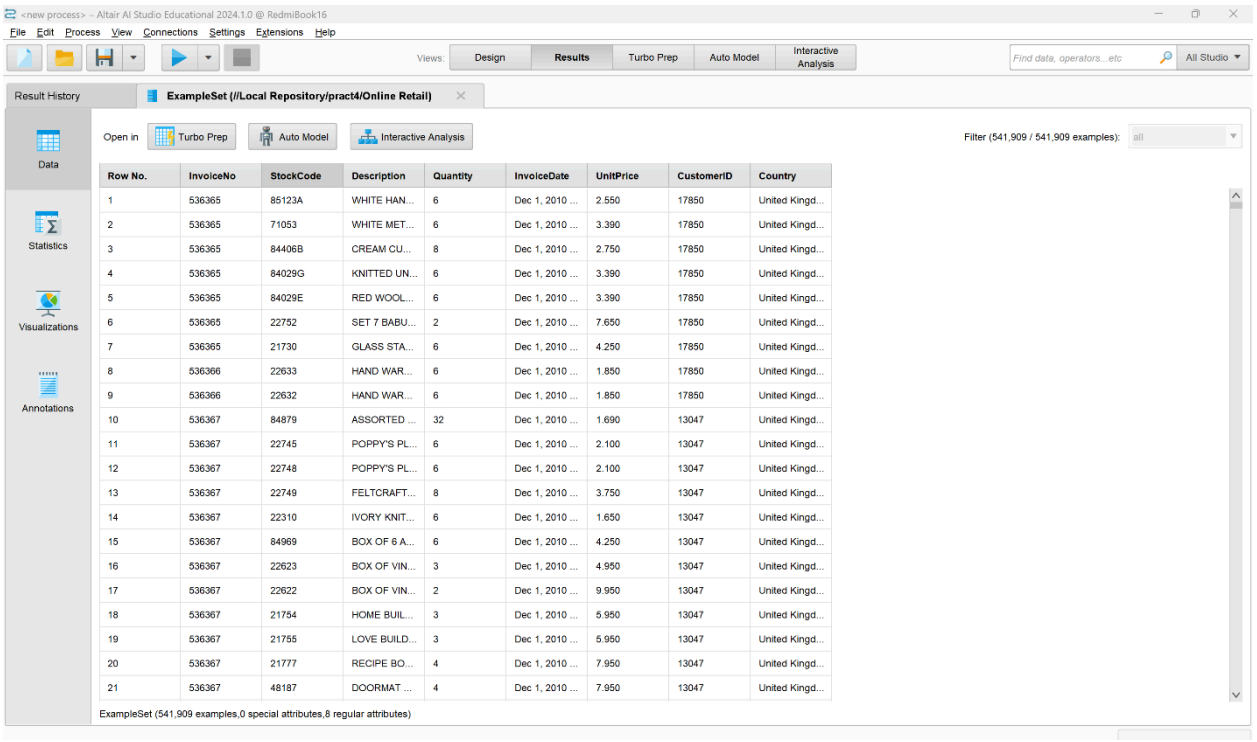


Рисунок 3.2 – выгруженные данные

Фильтрация данных:

Для начала мы удалим из исходного набора записей те данные, которые не могут быть использованы для анализа. Для этого мы добавим в наш проект оператор **"Filter Examples"**.

В число данных, которые невозможно использовать вошли транзакции с отрицательными значениями в столбце Quantity, что указывает на возвраты. Настройка фильтра представлена на рисунке 3.3.

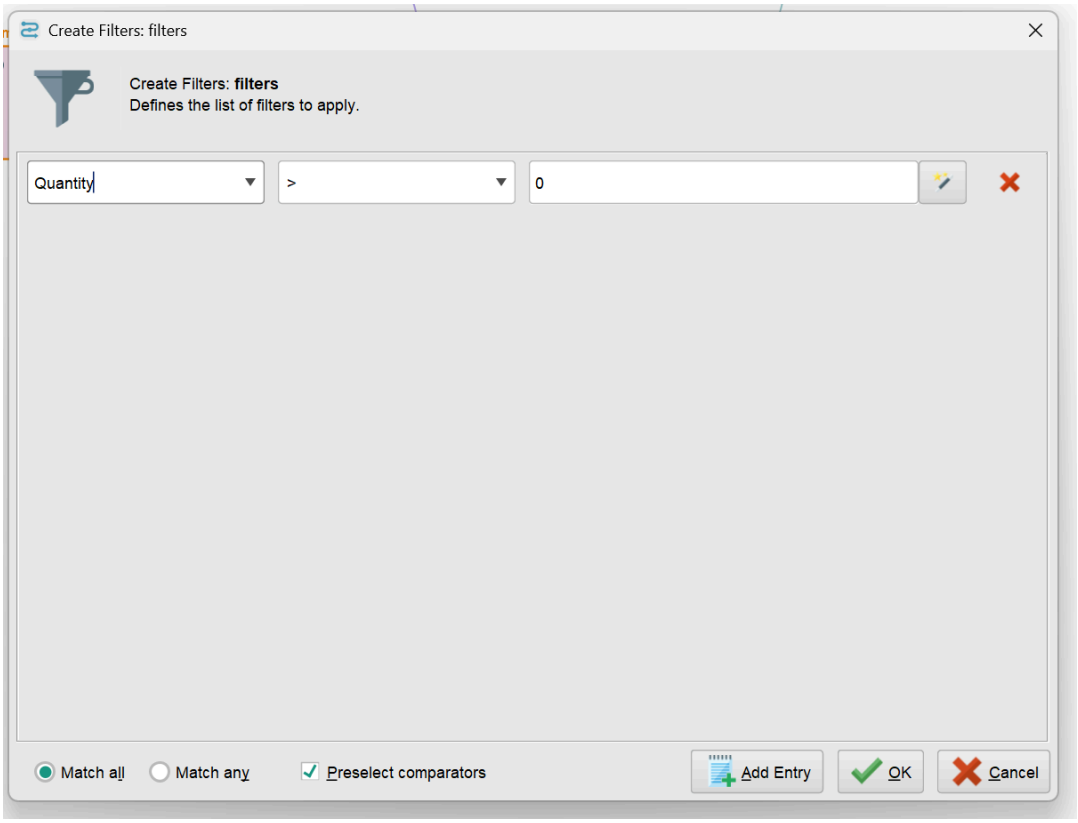


Рисунок 3.3 – настройки для Filter Examples

Агрегация данных:

Для дальнейшего анализа необходимо агрегировать данные, чтобы объединить записи, относящиеся к одной и той же транзакции. Для выполнения данной задачи в проект был добавлен оператор **"Aggregate"**, позволяющий агрегировать данные на основе выбранных атрибутов.

Обязательно необходимо объединить Aggragate с Filter Examples, который в свою очередь объединён с самими данными рисунок 3.4

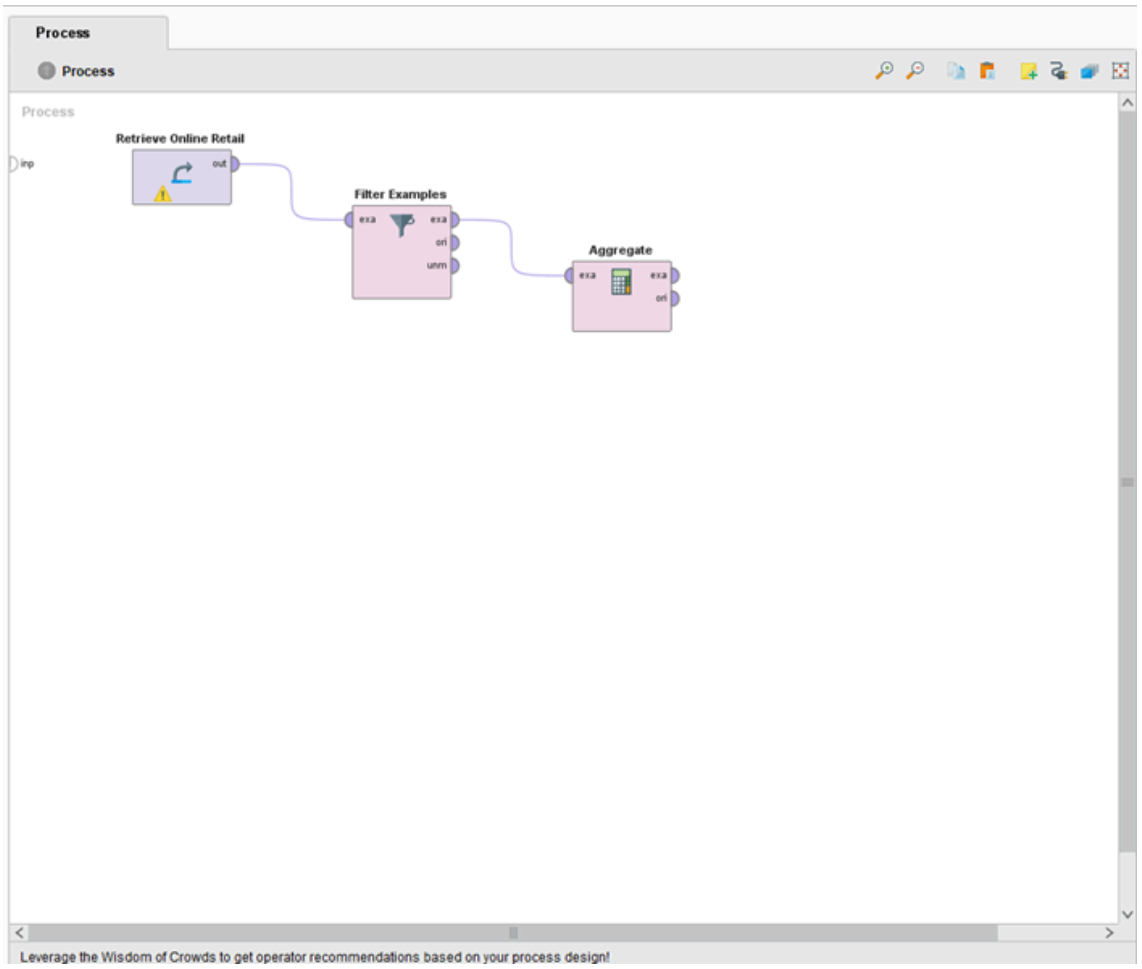


Рисунок 3.4 - Вид схемы с оператором Aggragate

В параметрах агрегатора в разделе "group by attributes" была указана колонка **InvoiceNo**, которая идентифицирует каждую уникальную транзакцию. Это позволило сгруппировать все записи, относящиеся к одному и тому же заказу.

В разделе "aggregation attributes" была выбрана колонка **Description**, содержащая наименования товаров. Для данной колонки была установлена функция **concatenation**, что позволило объединить все наименования товаров из одной транзакции в одну строку, разделённую символами.

В результате выполнения агрегирования каждый заказ представлен одной строкой с объединённым списком товаров. Это позволяет сократить объём данных и сделать их более наглядными для последующего анализа и построения ассоциативных правил.

Настройки данного блока представлены на рисунках. На рисунке 3.5 показаны параметры группировки, а на рисунке 3.6 - параметры агрегации.

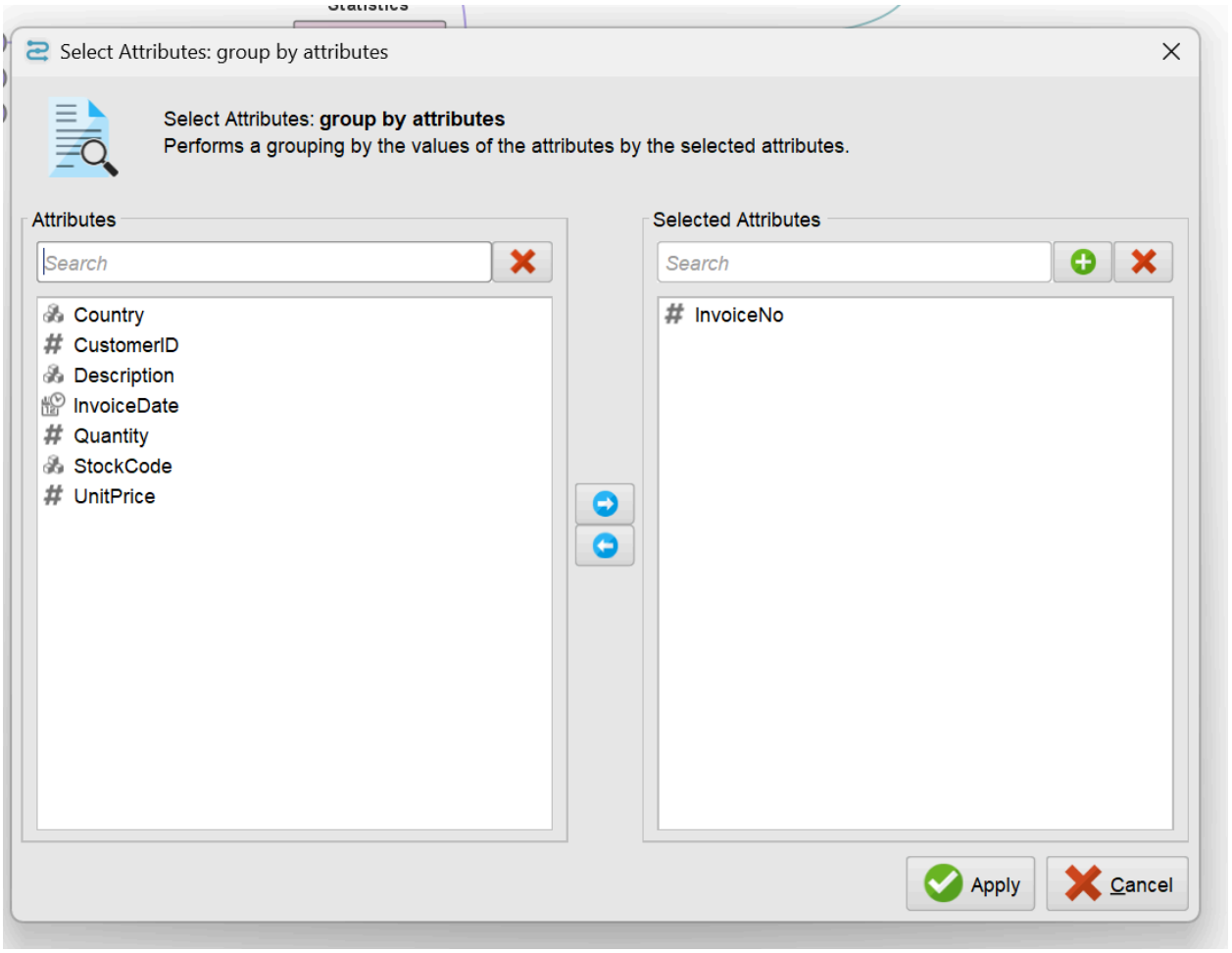


Рисунок 3.5 – настройки для группировки

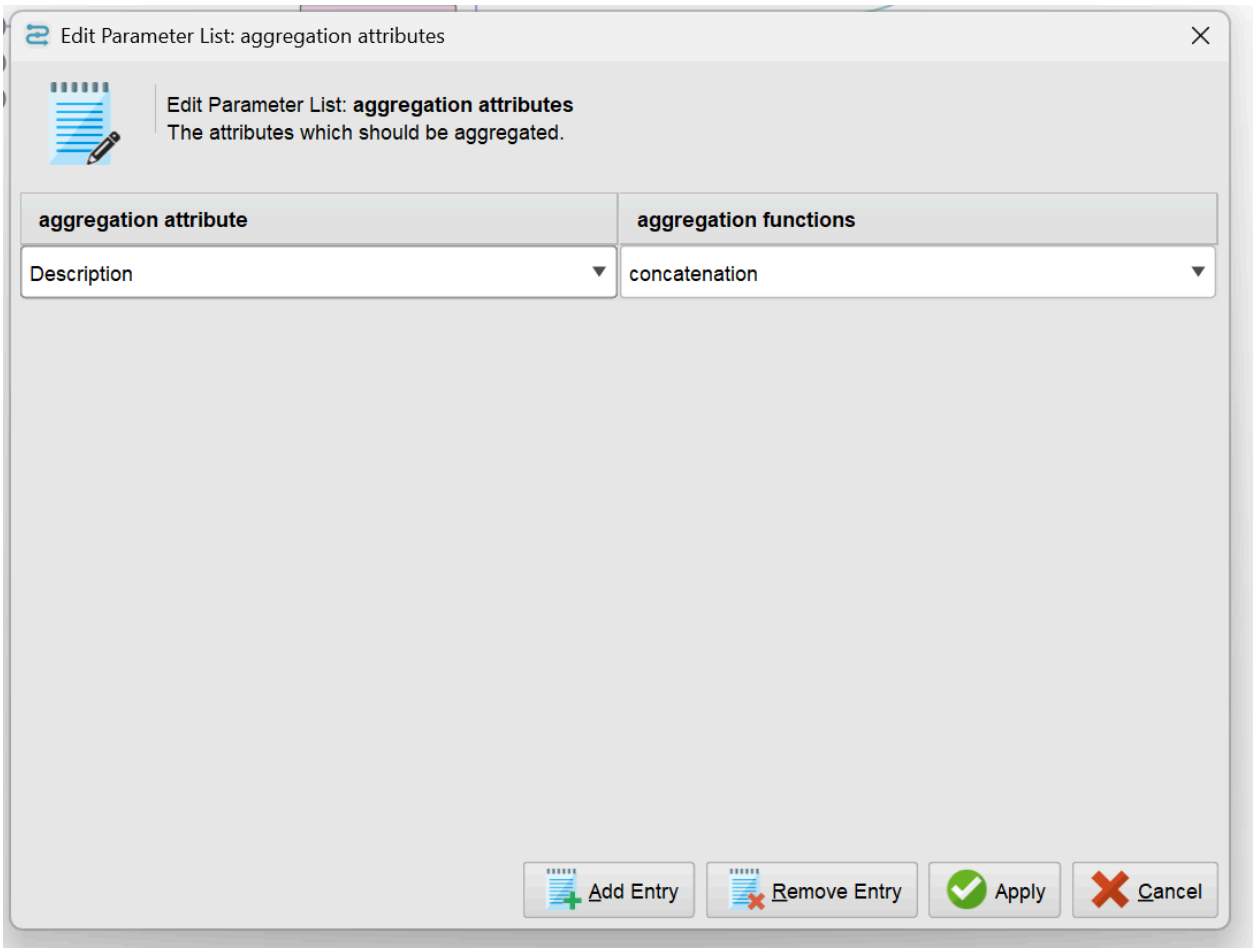


Рисунок 3.6 – настройки для агрегации

Выбор атрибутов:

Для подготовки данных к дальнейшей обработке и построению ассоциативных правил был добавлен оператор **"Select Attributes"**, предназначенный для отбора только необходимых столбцов из исходного набора данных. Также соединяем поле Aggregate рисунок 3.7.

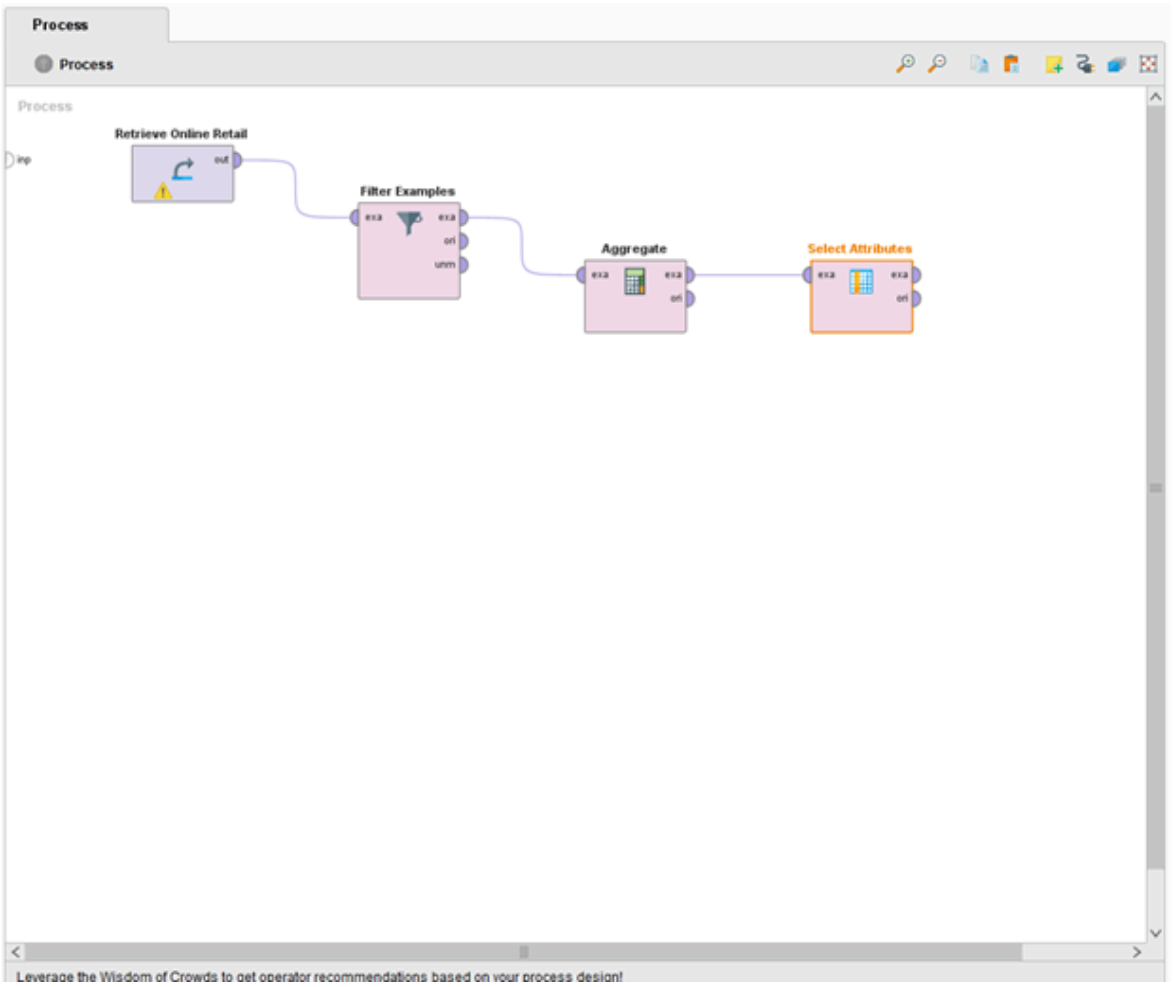


Рисунок 3.7 – схема подключения

В параметрах блока в разделе **"select subset"** были выбраны два атрибута: **InvoiceNo** (уникальный идентификатор каждой транзакции) и **concat (Description)** (объединённое описание товаров, полученное на предыдущем этапе агрегирования).

Столбец concat(Description) переносится в раздел Selected Attributes. Тип фильтрации был настроен как **"include attributes"**, что позволило оставить в наборе данных только указанные атрибуты.

После выполнения блока **Select Attributes** в данных остаются только два столбца. Настройки блока представлены на рисунке 3.8.

Далее получаем результат работы конструкции, объединяем выход exa к res и запускаем. Результаты применения блока показаны на рисунке 3.9

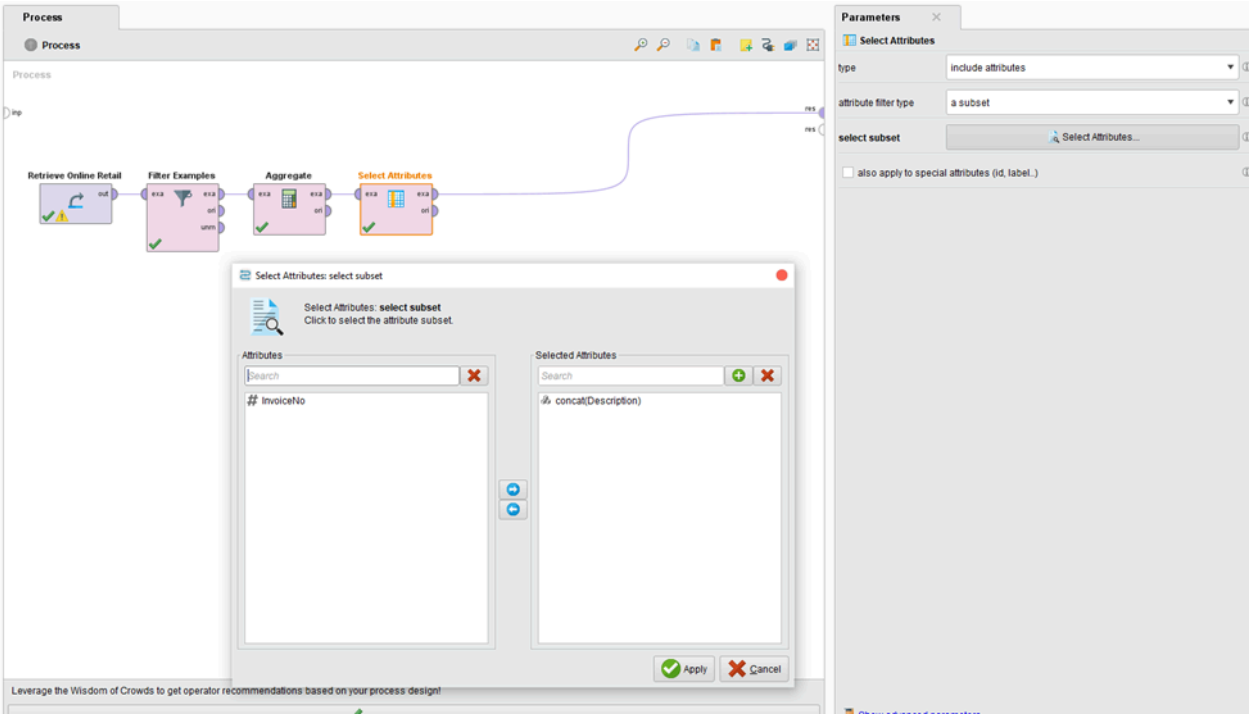


Рисунок 3.8 – настройки для Select Attributes

Row No.	concat(Description)
1	WHITE HANGING HEART T-LIGHT HOLDER|WHITE METAL LANTERN|CREAM CUPID HEARTS COAT HANGER|KNITTED UNION FLAG HOT WATER BOT...
2	HAND WARMER UNION JACK|HAND WARMER RED POLKA DOT
3	ASSORTED COLOUR BIRD ORNAMENT|POPPY'S PLAYHOUSE BEDROOM |POPPY'S PLAYHOUSE KITCHEN|FELTCRAFT PRINCESS CHARLOTTE DOLL...
4	JAM MAKING SET WITH JARS|RED COAT RACK PARIS FASHION|YELLOW COAT RACK PARIS FASHION|BLUE COAT RACK PARIS FASHION
5	BATH BUILDING BLOCK WORD
6	ALARM CLOCK BAKELIKE PINK|ALARM CLOCK BAKELIKE RED |ALARM CLOCK BAKELIKE GREEN|PANDA AND BUNNIES STICKER SHEET|STARS GIFT ...
7	PAPER CHAIN KIT 50'S CHRISTMAS
8	HAND WARMER RED POLKA DOT|HAND WARMER UNION JACK
9	WHITE HANGING HEART T-LIGHT HOLDER|WHITE METAL LANTERN|CREAM CUPID HEARTS COAT HANGER|EDWARDIAN PARASOL RED|RETRO COF...
10	VICTORIAN SEWING BOX LARGE
11	WHITE HANGING HEART T-LIGHT HOLDER|WHITE METAL LANTERN|CREAM CUPID HEARTS COAT HANGER|EDWARDIAN PARASOL RED|RETRO COF...
12	HOT WATER BOTTLE TEA AND SYMPATHY|RED HANGING HEART T-LIGHT HOLDER
13	HAND WARMER RED POLKA DOT|HAND WARMER UNION JACK
14	JUMBO BAG PINK POLKADOT|JUMBO BAG BAROQUE BLACK WHITE|JUMBO BAG CHARLIE AND LOLA TOYS|STRAWBERRY CHARLOTTE BAG|RED 3 PI...
15	JAM MAKING SET PRINTED
16	RETROSPOT TEA SET CERAMIC 11 PC |GIRLY PINK TOOL SET|JUMBO SHOPPER VINTAGE RED PAISLEY|AIRLINE LOUNGE,METAL SIGN|WHITE SPOT...
17	INFLATABLE POLITICAL GLOBE |VINTAGE SNAKES & LADDERS|CHOCOLATE CALCULATOR|JUMBO SHOPPER VINTAGE RED PAISLEY|RECYCLING BA...
18	WOOD BLACK BOARD ANT WHITE FINISH|COLOUR GLASS T-LIGHT HOLDER HANGING|HANGING METAL HEART LANTERN|HANGING MEDINA LANTER...
19	SET 3 WICKER OVAL BASKETS W LIDS|JAM MAKING SET PRINTED|JAM MAKING SET WITH JARS|JUMBO BAG DOLLY GIRL DESIGN|TRADITIONAL CHRI...
20	WHITE WIRE EGG HOLDER|JUMBO BAG BAROQUE BLACK WHITE|JUMBO BAG RED RETROSPOT
21	CHILLI LIGHTS|LIGHT GARLAND BUTTERFILES PINK|WOODEN OWLS LIGHT GARLAND |FAIRY TALE COTTAGE NIGHTLIGHT|RED TOADSTOOL LED NI...
22	HOME BUILDING BLOCK WORD|LOVE BUILDING BLOCK WORD|DOORMAT FANCY FONT HOME SWEET HOME|HOME SMALL WOOD LETTERS|GINGHA...
23	CHRISTMAS LIGHTS 10 REINDEER|VINTAGE UNION JACK CUSHION COVER|VINTAGE HEADS AND TAILS CARD GAME |SET OF 3 COLOURED FLYING ...
24	CHRISTMAS LIGHTS 10 REINDEER|JAM MAKING SET WITH JARS|JAM MAKING SET PRINTED|JAM JAR WITH PINK LID|JAM JAR WITH GREEN LID|ROSE...
25	3 STRIPEY MICE FELTCRAFT|SET OF 6 SOLDIER SKITTLES|TRADITIONAL WOODEN SKIPPING ROPE|WOODEN BOX OF DOMINOES|RUSTIC SEVENT...
26	RETROSPOT LAMP
27	FANCY FONT BIRTHDAY CARD, |HAND WARMER UNION JACK|HAND WARMER SCOTTY DOG DESIGN|HAND WARMER OWL DESIGN|HAND WARMER R...
28	BLACK HEART CARD HOLDER|ASSORTED COLOUR BIRD ORNAMENT|PACK OF 60 PINK PAISLEY CAKE CASES|60 TEATIME FAIRY CAKE CASES|PACK ...
29	WHITE HANGING HEART T-LIGHT HOLDER|WHITE METAL LANTERN|CREAM CUPID HEARTS COAT HANGER|EDWARDIAN PARASOL BLACK|EDWARDI...
30	SET OF 3 BLACK FLYING DUCKS|SET OF 3 COLOURED FLYING DUCKS
31	PACK OF 12 RED RETROSPOT TISSUES |RED RETROSPOT MUG|BABUSHKA LIGHTS STRING OF 10|PIGGY BANK RETROSPOT |SET 7 BABUSHKA NE...
32	HAND WARMER RED POLKA DOT|HAND WARMER UNION JACK
33	HOMEMADE JAM SCENTED CANDLES
34	BIRD HOUSE HOT WATER BOTTLE|BOUDOIR SQUARE TISSUE BOX|SKULLS SQUARE TISSUE BOX|PHOTO FRAME CORNICE|SILK PURSE BABUSHKA ...
35	PAPER CHAIN KIT 50'S CHRISTMAS |PAPER CHAIN KIT VINTAGE CHRISTMAS|HOT WATER BOTTLE BABUSHKA
36	HAND WARMER BIRD DESIGN|POSTAGE
37	HEART IVORY TRELLIS SMALL|CLEAR DRAWER KNOB ACRYLIC EDWARDIAN|PINK DRAWER KNOB ACRYLIC EDWARDIAN|GREEN DRAWER KNOB AC...

ExampleSet (20.726 examples,0 special attributes,1 regular attribute)

Рисунок 3.9 – результаты применения Select Attributes

FP-Growth:

FP-Growth (Frequent Pattern Growth) - это алгоритм для нахождения часто встречающихся наборов элементов в транзакционных данных. В нашей лабораторной работе этот оператор используется для выявления частых наборов товаров, которые покупатели заказывают вместе.

Параметр **"input format"** установлен как **"item list in a column"**, что означает, что данные представляют собой список элементов (товаров), сгруппированных по транзакциям в одной колонке.

Разделитель элементов (Item separators):

Указан разделитель **"|",** который используется для разделения товаров в объединённых строках. Это важно, так как описание товаров после предварительной обработки имеет именно такой формат.

Тримминг названий товаров (Trim item names):

Опция **"trim item names"** активирована, что позволяет удалить лишние пробелы в названиях товаров, предотвращая ошибки при обработке данных.

Минимальная поддержка (Min support):

Установлено значение **0.02**, что означает, что будут учитываться только те наборы товаров, которые встречаются не менее чем в 2% всех транзакций. Этот параметр позволяет исключить редкие сочетания, которые не имеют значимой ценности для анализа.

Минимальное количество элементов в наборе (Min items per itemset):

Значение **1** указывает, что алгоритм будет рассматривать наборы товаров, состоящие как минимум из одного элемента.

Максимальное количество элементов в наборе (Max items per itemset):

Значение **0** означает отсутствие ограничений на максимальное количество товаров в наборе.

Максимальное количество наборов (Max number of itemsets):

Параметр установлен на значение **1,000,000**, что позволяет обработать до миллиона наборов, если такие будут выявлены.

Автоматический выбор минимального количества наборов (Find min number of itemsets):

Активирована опция **"find min number of itemsets"**, что позволяет алгоритму автоматически определять оптимальное минимальное количество наборов для анализа.

Минимальное количество наборов (Min number of itemsets):

Указано значение **100**, что гарантирует выбор хотя бы 100 часто встречающихся наборов.

После выполнения блока FP-Growth мы получаем список часто встречающихся наборов товаров с указанием их поддержки. Эти данные являются основой для создания ассоциативных правил в следующем блоке. Полный список настроек для блока представлен на рисунке 3.10.

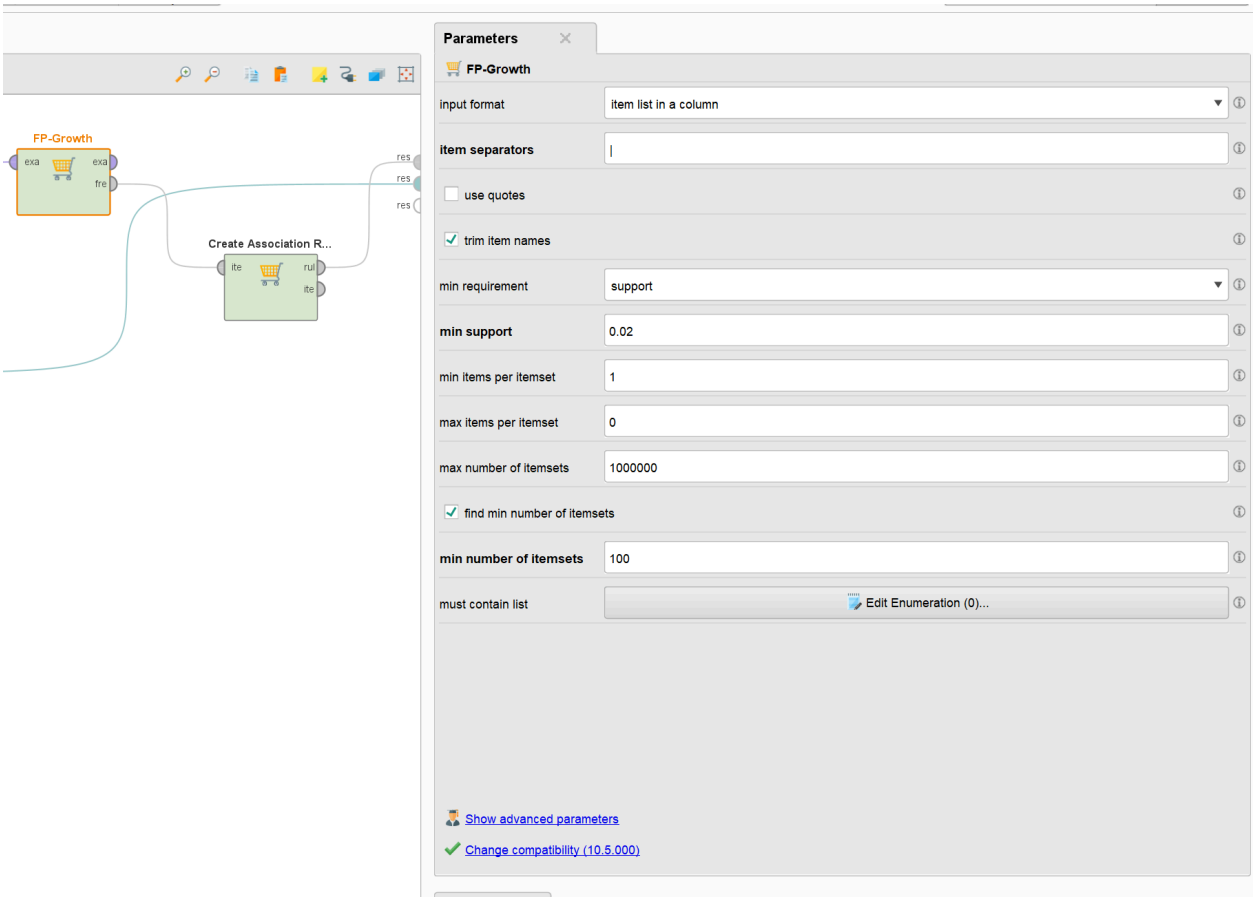


Рисунок 3.10 – Настройка FP-Growth

Создание ассоциативных правил:

Блок **"Create Association Rules"** используется для генерации ассоциативных правил на основе частых наборов, выявленных оператором **FP-Growth**. Ассоциативные правила помогают определить взаимосвязи между товарами, которые часто покупаются вместе.

Установлено значение **confidence**, что означает использование метрики доверия для фильтрации и ранжирования правил. Доверие показывает, с какой вероятностью товар B покупается вместе с товаром A.

Указано значение **0.5**, что означает, что в результирующий набор войдут только те правила, для которых вероятность сопутствующей покупки превышает 50%. Это позволяет отфильтровать правила с низким уровнем достоверности и сосредоточиться на значимых связях.

После выполнения блока **Create Association Rules** мы получаем список ассоциативных правил.

Настройки блока показаны на рисунке 3.11. Финальный вид схемы представлен на рисунке 3.12. Набор полученных правил представлен на рисунке 3.13. Его можно найти на вкладке **"Description"** после успешного применения блока.

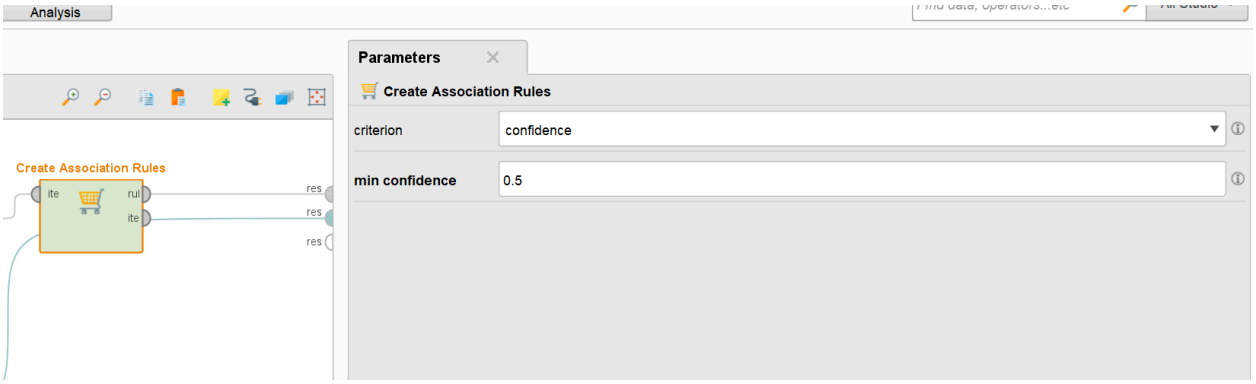


Рисунок 3.11 – Настройка блока Create Association Rules

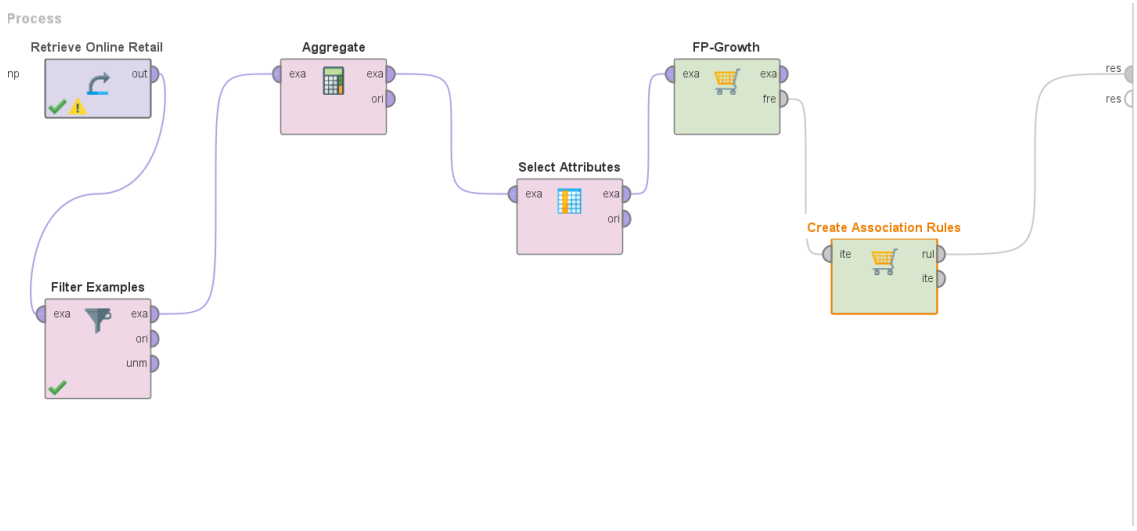


Рисунок 3.12 – Финальный вид схемы

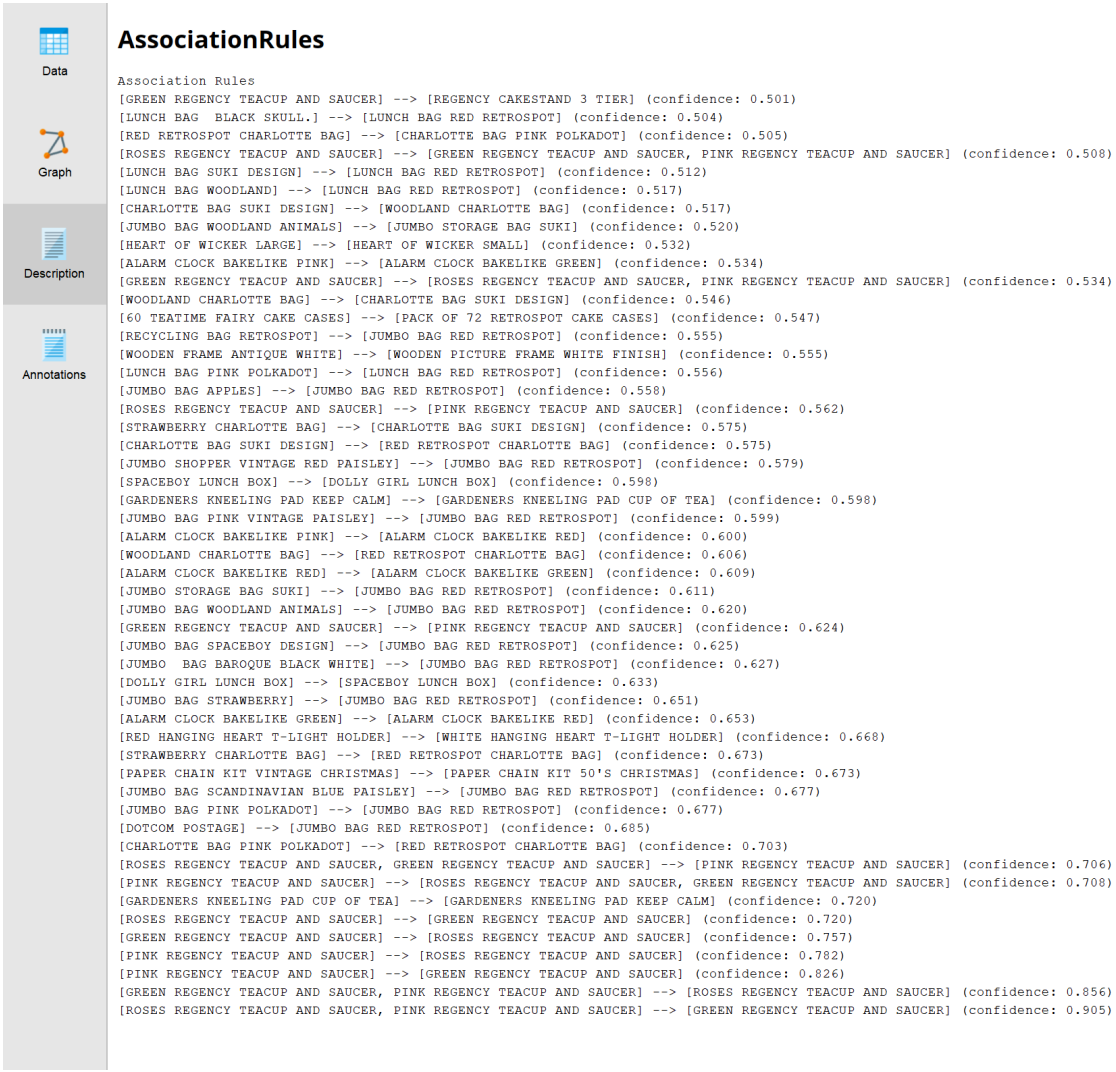


Рисунок 3.13 – Полученные ассоциативные правила

На рисунке 3.11 представлен раздел Association Rules, где показаны выявленные ассоциативные правила. Эти правила описывают зависимости между товарами, которые чаще всего приобретаются совместно.

Для каждого правила приведён показатель уверенности, отражающий вероятность покупки товаров из следствия, если были приобретены товары из предпосылки.

Рассмотрим правило [ROSES REGENCY TEACUP AND SAUCER] --> [PINK REGENCY TEACUP AND SAUCER] (уверенность 56,2%) демонстрирует, что покупатели часто дополняют чайные сервизы чашками другого цвета, что указывает на их интерес к составлению гармоничных наборов.

В результате мы приходим к очевидным выводам - товары с близкими характеристиками (например, из одной линейки или с похожими стилями) показывают

высокую вероятность совместной покупки. Это важно для оптимизации ассортимента и выкладки. Многие правила включают товары с одинаковыми основными характеристиками, такими как цвет, форма или функциональность.

Таким образом ассоциативный анализ может использоваться для разработки маркетинговых стратегий, оптимизации расположения товаров в магазине, выявления скрытых предпочтений покупателей, которые не очевидны на первый взгляд.

Визуализация результатов с помощью графа

После создания ассоциативных правил, мы можем использовать функцию **"Graph"**, которая предоставляет возможность визуального представления связей между элементами в данных. Данный инструмент позволяет лучше понять, какие связи преобладают, а также наглядно представить зависимость между товарами и частоту их совместного появления в транзакциях.

Каждый узел на графе представляет товар, а связи между узлами обозначают ассоциативные правила, созданные ранее. На рисунке 3.12 представлен первоначальный вид графа. Для уменьшения количества отображаемых данных был установлен минимальный уровень доверия на уровне близком к 1.0. Это позволило исключить менее значимые правила, сосредоточив внимание на более надежных связях рисунок 2.15.

Рисунок 3.15 – первоначальный вид графа

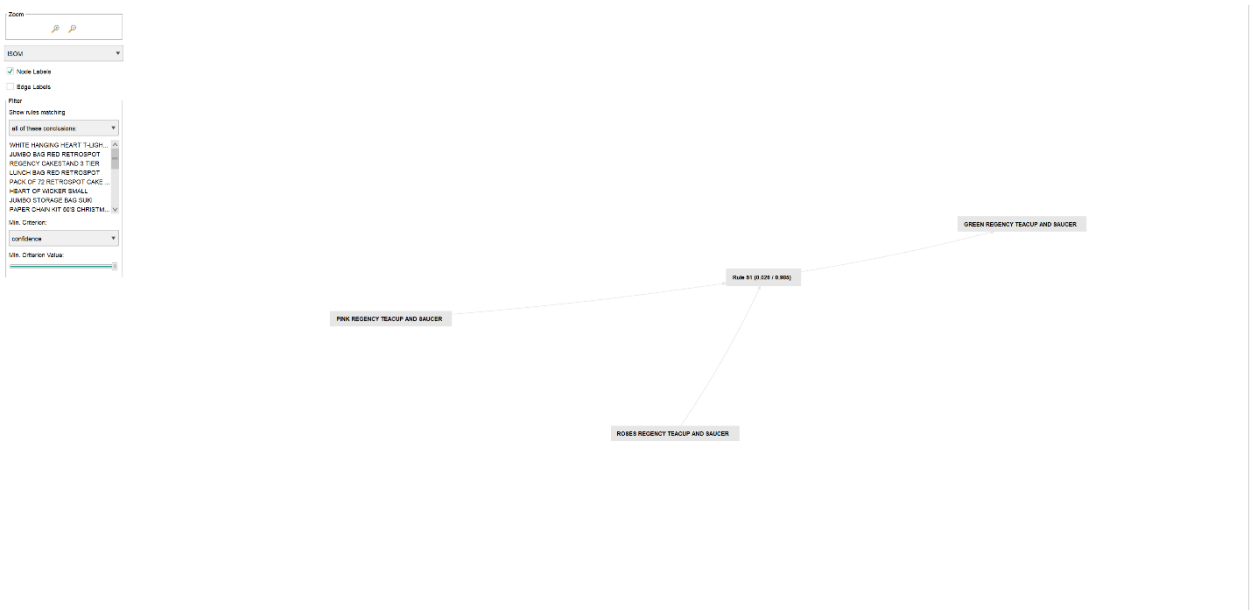


Рисунок 3.13 – граф с параметрами confidence близкими к 1.0

Исходя из полученных результатов, можно сделать ряд выводов. Товары, такие как JUMBO BAG RED RETROSPOT и PINK REGENCY TEACUP AND SAUCER, показали наибольшее количество взаимосвязей с другими продуктами. Это указывает на их высокую популярность и универсальность в покупательских предпочтениях. Эти товары могут быть основой для создания рекомендаций в маркетинговых кампаниях.

На графе выделяются группы товаров, которые покупаются вместе. Например, различные виды чайных чашек (GREEN REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER) часто связаны друг с другом. Это может быть полезно для формирования товарных комплектов или акций.

Некоторые товары имеют небольшое количество связей, но при этом демонстрируют высокую надежность ассоциативных правил. Это говорит о том, что они приобретаются в специфических условиях и могут быть интересны для целевой аудитории.

Граф также показал изолированные группы товаров, которые практически не пересекаются с другими кластерами. Это может свидетельствовать о существовании отдельных сегментов покупателей с уникальными предпочтениями.

Использование блока Statistics:

Финальным блоком, рассмотренным в рамках данной лабораторной работы, станет блок **"Statistics".** Он необходим для получения общей информации о наборе

данных и ключевых статистических показателей. Это позволяет не только лучше понять структуру данных, но и выявить основные характеристики, которые могут быть полезны для анализа. Схема подключения к уже существующим блокам представлена на рисунке 3.16. А часть полученных результатов на рисунках 3.17, 3.18 и 3.19

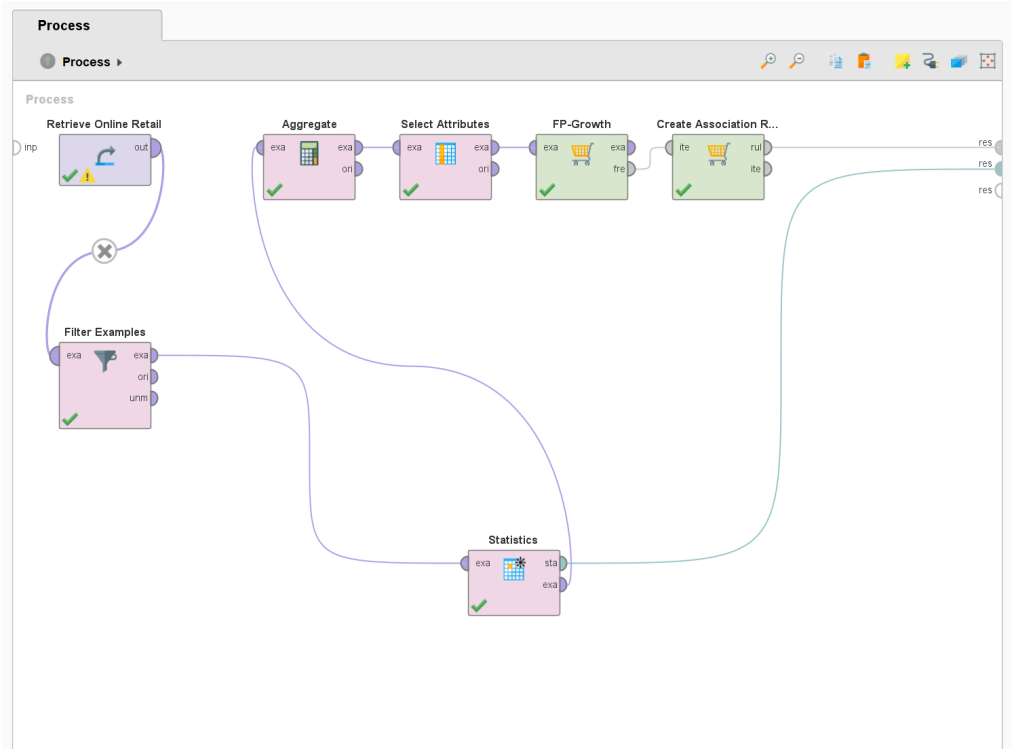


Рисунок 3.16 – схема с подключенным блоком Statistics

Рисунок 3.17 – Статистика для атрибута "Description"

Рисунок 3.18 – Статистика для атрибута "StockCode"

Рисунок 3.19 – Статистика для атрибута "Country"

На рисунках представлена статистика по основным атрибутам набора данных: "Description", "StockCode" и "Country". Каждый из этих атрибутов был подробно проанализирован для выявления ключевых закономерностей и распределений.

Анализ атрибута "Description", который представляет наименования товаров. Как нам уже удалось выяснить - наиболее частые товары: WHITE HANGING HEART T-LIGHT HOLDER, JUMBO BAG RED RETROSPOT и REGENCY CAKESTAND 3.

Общее количество уникальных наименований: 4,077, что указывает на большое разнообразие ассортимента.

На графике, где представлен анализ кодов товаров ("StockCode"). Мы получаем коды 85123A и 85099B, встречаются наиболее часто, и соответствуют HANGING HEART T-LIGHT HOLDER и JUMBO BAG RED RETROSPOT. Хотя уникальные значения: 3,941 немного меньше, чем для "Description", это логично, так как некоторые товары могут иметь одинаковые коды. А пропущенные значения: отсутствуют, что делает этот атрибут полностью готовым для анализа.

Третий график иллюстрирует распределение транзакций по странам: Великобритания (United Kingdom) занимает 91.95% всех записей, что подчёркивает доминирующую роль этой страны в наборе данных.

4. Приобретаемые навыки

1. Работа с интерфейсом RapidMiner Studio для анализа данных и построения процессов.
2. Умение загружать, очищать и подготавливать данные для построения ассоциативных правил.
3. Применение алгоритмов *FP-Growth* и *Create Association Rules* для анализа паттернов.
4. Построение визуализаций в виде графов для наглядного отображения связей между элементами.
5. Использование операторов для проведения статистического анализа данных.
6. Анализ результатов статистики, выявление наиболее популярных товаров и их распределения.
7. Развитие навыков подготовки отчетов и визуализации данных для представления результатов анализа.

5. Обобщенная задача для индивидуального варианта

Цель работы – провести анализ ассоциативных правил на предоставленном наборе данных. В вашей работе должны быть реализованы следующие этапы:

1) Загрузка и предобработка данных

- Импортируйте таблицу с транзакциями, содержащую как минимум два столбца: идентификатор транзакции и список элементов (товаров, событий и т.п.).
- Отфильтруйте ненужные записи (возвраты, пустые транзакции), удалите лишние атрибуты и преобразуйте формат данных так, чтобы в одной колонке для каждой транзакции был перечислен через разделитель полный список элементов.

2) Выявление частых наборов элементов

- Примените алгоритм FP-Growth или Apriori для поиска частых паттернов.

- Настройте пороги min support и max items per itemset, чтобы отобрать устойчивые и информативные сочетания.

3) Построение ассоциативных правил

- Используйте оператор Create Association Rules для генерации правил на основе выявленных частых наборов.
- Установите порог min confidence и при необходимости порог min lift, чтобы отсеять нерелевантные или слишком общие правила.

4) Анализ и визуализация

- Отберите топ-N правил по разным метрикам (confidence, lift, support) и проанализируйте их бизнес-значимость.
- Постройте граф связей (Graph) или тепловую матрицу, чтобы наглядно отобразить основные закономерности и кластеры товаров/событий, часто встречающихся вместе.

5) Выводы

- Сформулируйте ключевые выводы: какие сочетания элементов наиболее значимы и почему.
- Дайте практические рекомендации (кросс-продажи, компоновка витрин, персональные предложения), основанные на полученных правилах.

6. Распределение вариантов