УДК 004

Разведочный анализ данных о продажах товаров с помощью программного пакета визуального программирования Orange

Голубева Евгения Павловна
Приамурский государственный университет имени Шолом-Алейхема
Студент

Аннотация
Цель данной статьи – выполнить разведочный анализ данных о продажах товаров в период 2020-2022 г. Для разведочного анализа был использован программный пакет визуального программирования на основе компонентов для визуализации данных Orange и данные о продажах. С помощью средств визуализации Orange выполнили разведочный анализ данных о продажах и получили итоговую схему.
Ключевые слова: Orange, виджет, визуализация данных, разведочный анализ.

Exploratory Data Analysis in Orange

Golubeva Evgeniya Pavlovna
Sholom-Aleichem Priamursky State University
Student

Abstract
The purpose of this article is to perform an exploratory analysis of data on sales of goods in the period 2020-2022. For exploratory analysis, a visual programming software package based on orange data visualization components and sales data was used. With the help of Orange visualization tools, we performed an exploratory analysis of sales data and obtained the final scheme.
Keywords: Orange, widget, data visualization, exploration analysis

1 Введение

1.1 Актуальность

Информация играет очень важную роль в современной жизни. На сегодняшний день объемы информации колоссальны и продолжают расти.

Разведочный анализ данных позволяет изучить данные до самых глубин для получения из них практической информации. Разведочный анализ включает в себя анализ и обобщение массивных наборов данных, часто в форме диаграмм и графиков.

Программа Orange благодаря своим функциям позволяет провести разведочный анализ данных.

1.2 Обзор исследований

В. Е. Максимов, К. М. Резникова и Д. А. Попов исследовали применение программы Orange Data Mining для автоматизированного интеллектуального

анализа данных морского флота [1]. В данной статье дали подробную характеристику каждому инструменту Data Mining Ш. Е. Омарова и А. М. Медеубаева [2]. А. Т. Маматкасымова и Н. А. Кульматова показали принципы работы системы визуального программирования Orange при работе с объемными данными [3]. В статье рассматривали исследование программного обеспечения Data Mining Ю. С. Кривенко, А.Т. Минасян и А.О. Разиньков [4]. С.В. Пальмов и А.А. Жуйкова в статье описали функционал аналитического пакета Orange, предназначенный для поиска часто встречающихся наборов элементов и ассоциативных правил [5]. Изучили популярные методы анализа данных и разведочный анализ данных и Data Mining А. И. Токарев и А.Н. Брякин [6]. С. С. Мастевой и А.Н. Петрова статью посвятили краткому обзору методов Data Mining в период информационной эры [7].

1.3 Цель исследования

Цель исследования - выполнить разведочный анализ данных с помощью программы Orange.

2 Материалы и методы

Для разведочного анализа используется программа Orange. Работа будет происходить на готовых данных Данные.xlsx, скачать которые можно по ссылке:
https://docs.google.com/spreadsheets/d/1C477TrbfzWdGK1MYv6MDgwbpPd1Dnn7H/edit?usp=share_link&ouid=104272149632818699735&rtpof=true&sd=true

3 Результаты и обсуждения

Перед началом работы требуется установить Orange с официального сайта и установить.

1) Создадим новый файл (рис.1).

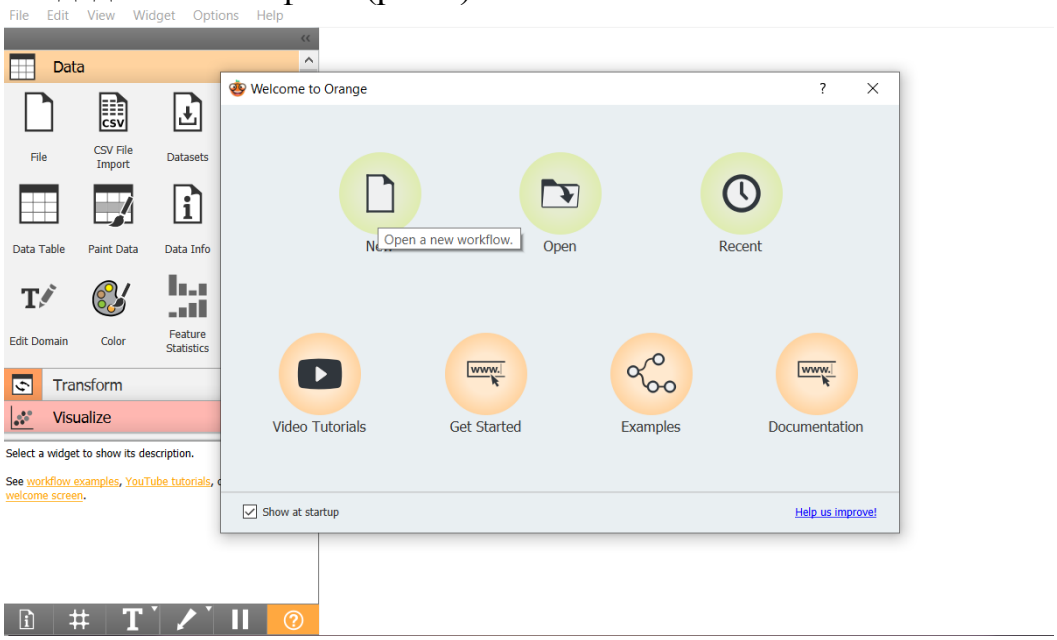


Рис.1. Создание нового файла

2) Добавляем виджет File на холст (Рис.2).

Виджет File считывает файл входных данных (таблицу данных с экземплярами данных) и отправляет набор данных в выходной канал. История последних открытых файлов хранится в виджете. Виджет также включает каталог с образцами наборов данных, которые поставляются с предустановленным с Orange.

Виджет считывает данные из Excel (.xlsx), простых файлов с разделителями табуляции (.txt), файлов, разделенных запятыми (.csv) или URL-адресов.

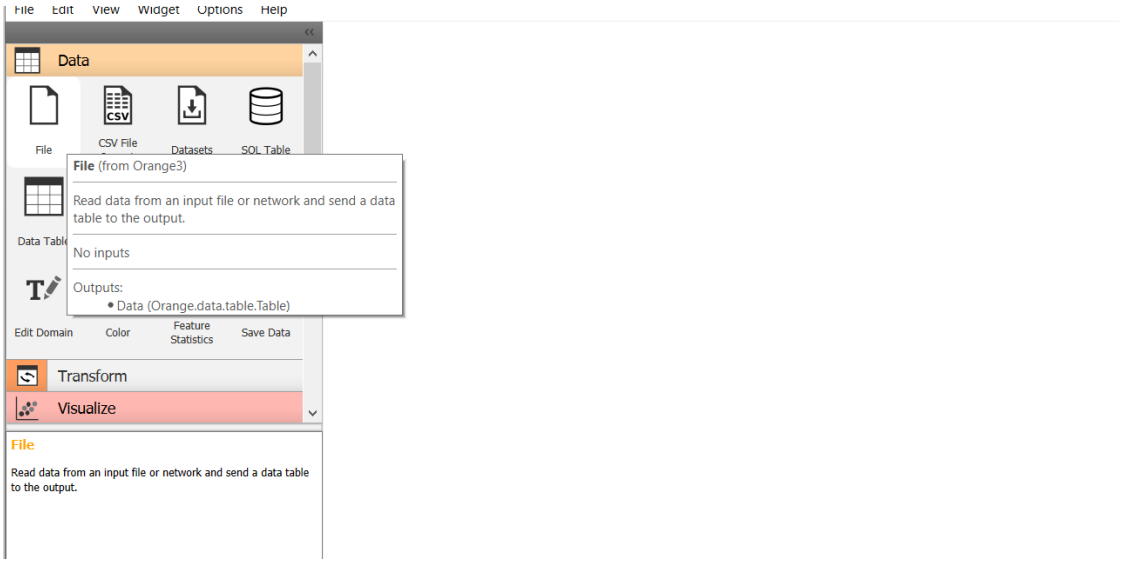


Рис.2. Добавление виджета File на холст

Есть 3 способа добавления виджета на холст:

1. Дважды щелкните на виджете.
2. Перетащите виджет на холст.
3. Щелкните правой кнопкой мыши на холсте для меню виджета.

3) Чтобы добавить файл необходимо открыть виджет file на холсте (Рис.3).

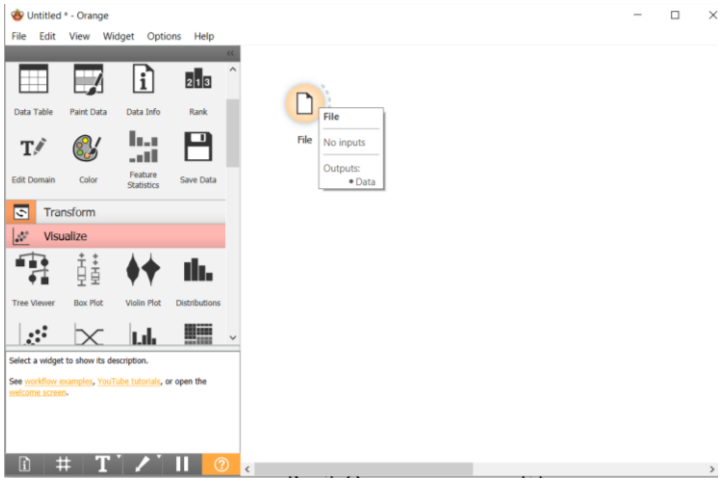


Рис.3. Открытие виджета File

4) Открылось диалоговое окно File (Рис.4).

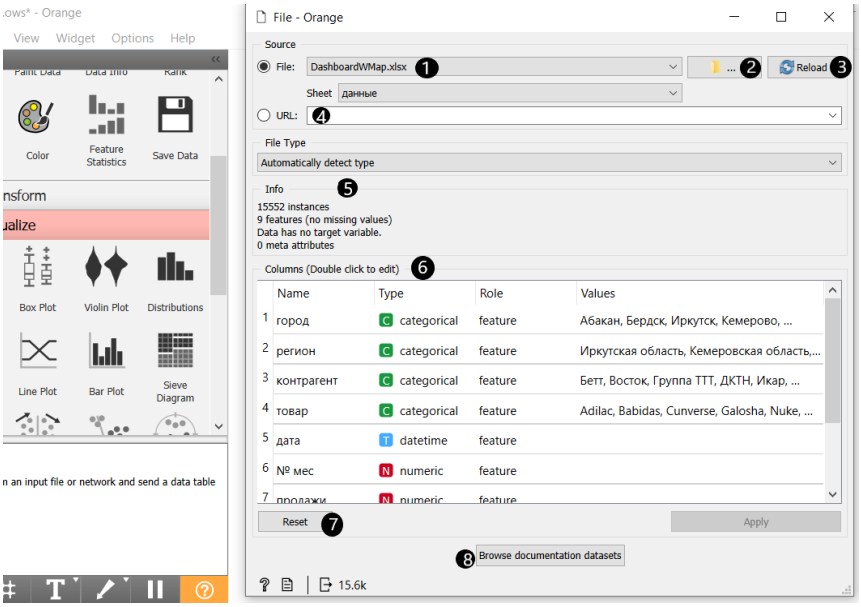


Рис.4. Диалоговое окно File

Описание диалогового окна File (Рис.4).

1. Просмотр ранее открытых файлов данных или загрузка любой из образцов.
2. Найти файл данных.
3. Перезагрузка выбранного в данный момент файл данных.
4. Вставка данных из URL-адресов, включая данные из Google Таблиц.
5. Информация о загруженном наборе данных: размер набора данных, количество и типы объектов данных.
6. Дополнительные сведения о функциях в наборе данных. Объекты можно редактировать, дважды щелкнув по ним. Пользователь может изменить имена атрибутов, выбрать тип переменной для каждого атрибута (Continuous, Nominal, String, Datetime) и выбрать способ дальнейшего определения атрибутов (как Features, Targets или Meta). Пользователь также может проигнорировать атрибут.
7. Сброс.
8. Просмотр наборов данных документации.

5) Добавляем файл Данные.xlsx (Рис.5).

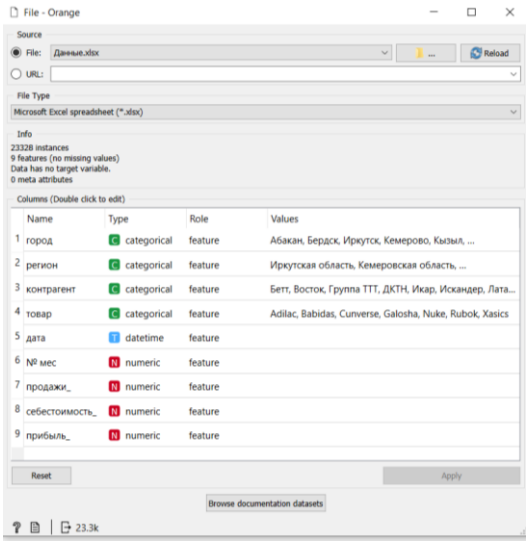


Рис.5. Добавление файла

6) Добавляем виджет Data Table на холст (Рис.6).

Data Table - получает один или несколько наборов данных в виде входных данных и представляет их в виде электронной таблицы. Экземпляры данных могут быть отсортированы по значениям атрибутов. Виджет также поддерживает ручной выбор экземпляров данных.

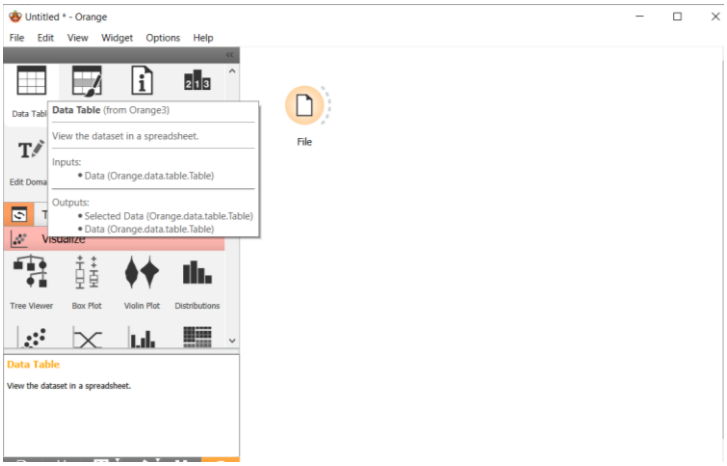


Рис.6. Добавление виджета Data Table

7) Чтобы посмотреть данные таблицы необходимо соединить два виджета на холсте File и Data Table (Рис.7).

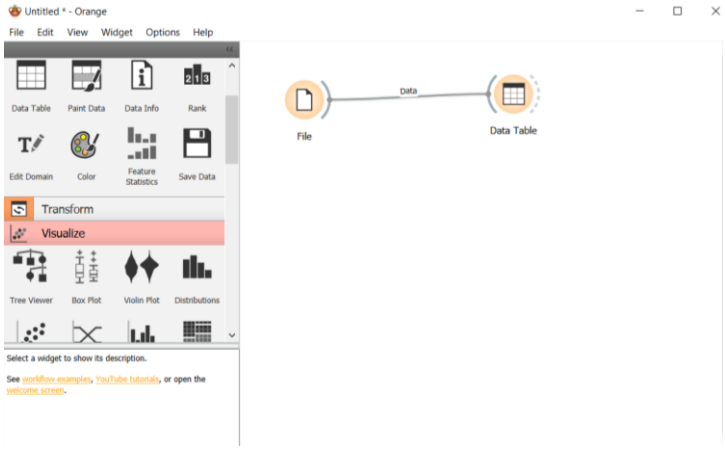
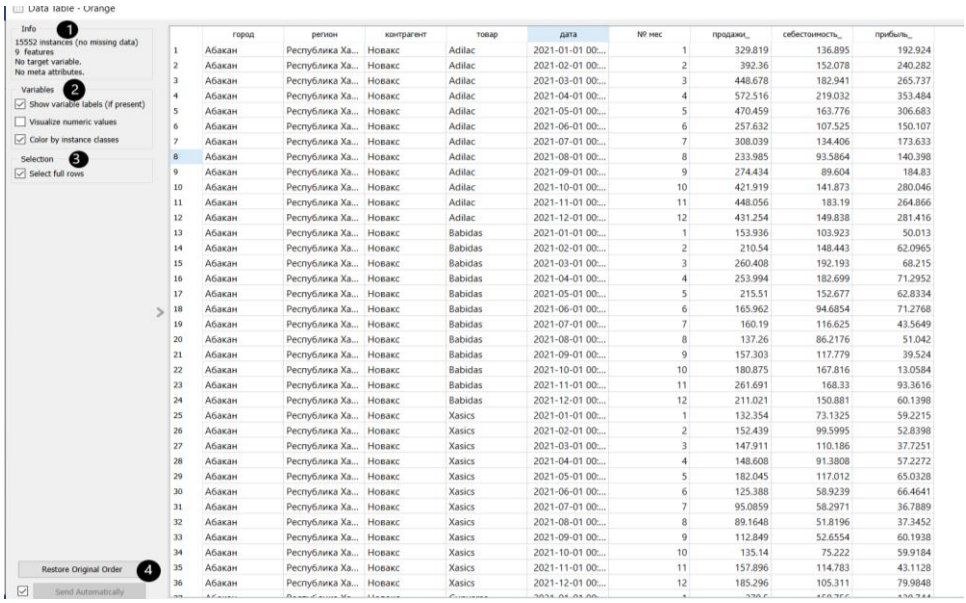


Рис.7. Соединение виджетов

8) Открываем виджет Data Table что бы просмотреть данные загруженной таблицы (Рис.8).



	город	регион	контрагент	товар	дата	№ мес	продажи_	себестоимость_	прибыль_
1	Абакан	Республика Ха...	Новакс	Adilac	2021-01-01 00:...	1	329.819	136.895	192.924
2	Абакан	Республика Ха...	Новакс	Adilac	2021-02-01 00:...	2	392.36	152.078	240.282
3	Абакан	Республика Ха...	Новакс	Adilac	2021-03-01 00:...	3	448.678	182.941	265.737
4	Абакан	Республика Ха...	Новакс	Adilac	2021-04-01 00:...	4	572.516	219.032	353.484
5	Абакан	Республика Ха...	Новакс	Adilac	2021-05-01 00:...	5	470.459	163.776	306.683
6	Абакан	Республика Ха...	Новакс	Adilac	2021-06-01 00:...	6	257.632	107.525	150.107
7	Абакан	Республика Ха...	Новакс	Adilac	2021-07-01 00:...	7	308.039	134.406	173.633
8	Абакан	Республика Ха...	Новакс	Adilac	2021-08-01 00:...	8	233.985	93.5864	140.398
9	Абакан	Республика Ха...	Новакс	Adilac	2021-09-01 00:...	9	274.434	89.604	184.83
10	Абакан	Республика Ха...	Новакс	Adilac	2021-10-01 00:...	10	421.919	141.873	280.046
11	Абакан	Республика Ха...	Новакс	Adilac	2021-11-01 00:...	11	448.056	183.19	264.866
12	Абакан	Республика Ха...	Новакс	Adilac	2021-12-01 00:...	12	431.254	149.838	281.416
13	Абакан	Республика Ха...	Новакс	Babidas	2021-01-01 00:...	1	153.936	103.923	50.013
14	Абакан	Республика Ха...	Новакс	Babidas	2021-02-01 00:...	2	210.54	148.443	62.0965
15	Абакан	Республика Ха...	Новакс	Babidas	2021-03-01 00:...	3	260.408	192.193	68.215
16	Абакан	Республика Ха...	Новакс	Babidas	2021-04-01 00:...	4	253.994	182.699	71.2952
17	Абакан	Республика Ха...	Новакс	Babidas	2021-05-01 00:...	5	215.51	152.677	62.8334
18	Абакан	Республика Ха...	Новакс	Babidas	2021-06-01 00:...	6	165.962	94.6854	71.2768
19	Абакан	Республика Ха...	Новакс	Babidas	2021-07-01 00:...	7	160.19	116.625	43.5649
20	Абакан	Республика Ха...	Новакс	Babidas	2021-08-01 00:...	8	137.26	86.2176	51.042
21	Абакан	Республика Ха...	Новакс	Babidas	2021-09-01 00:...	9	157.303	117.779	39.524
22	Абакан	Республика Ха...	Новакс	Babidas	2021-10-01 00:...	10	180.875	167.816	13.0584
23	Абакан	Республика Ха...	Новакс	Babidas	2021-11-01 00:...	11	261.691	168.33	93.3616
24	Абакан	Республика Ха...	Новакс	Babidas	2021-12-01 00:...	12	211.021	150.881	60.1398
25	Абакан	Республика Ха...	Новакс	Xasics	2021-01-01 00:...	1	132.354	73.1325	59.2215
26	Абакан	Республика Ха...	Новакс	Xasics	2021-02-01 00:...	2	152.439	99.5995	52.8398
27	Абакан	Республика Ха...	Новакс	Xasics	2021-03-01 00:...	3	147.911	110.186	37.7251
28	Абакан	Республика Ха...	Новакс	Xasics	2021-04-01 00:...	4	148.608	91.3808	57.2272
29	Абакан	Республика Ха...	Новакс	Xasics	2021-05-01 00:...	5	182.045	117.012	65.0328
30	Абакан	Республика Ха...	Новакс	Xasics	2021-06-01 00:...	6	125.388	58.9239	66.4641
31	Абакан	Республика Ха...	Новакс	Xasics	2021-07-01 00:...	7	95.0859	58.2971	36.7889
32	Абакан	Республика Ха...	Новакс	Xasics	2021-08-01 00:...	8	89.1648	51.8196	37.3452
33	Абакан	Республика Ха...	Новакс	Xasics	2021-09-01 00:...	9	112.849	52.6554	60.1938
34	Абакан	Республика Ха...	Новакс	Xasics	2021-10-01 00:...	10	135.14	75.222	59.9184
35	Абакан	Республика Ха...	Новакс	Xasics	2021-11-01 00:...	11	157.896	114.783	43.1128
36	Абакан	Республика Ха...	Новакс	Xasics	2021-12-01 00:...	12	185.296	105.311	79.9848

Рис.8. Диалоговое окно Data Table

Описание диалогового окна Data Table (Рис.8).

1. Сведения о текущем размере набора данных, количестве и типах атрибутов
2. Значения непрерывных атрибутов могут быть визуализированы с помощью баров; знамя может быть отнесен к различным классам.
3. Экземпляры данных (строки) могут быть выбраны и отправлены на выход виджета канал.
4. Используется кнопка Restore Original Order для изменения порядка экземпляров данных после сортировки на основе атрибутов.

9) Разведочный анализ данных (РАД) включает в себя визуализацию данных. Для визуализации данных будем использовать виджеты из раздела Visualize (Рис.9).

Рис.9. Раздел Visualize

10) Добавляем Box Plot из раздела Visualize (Рис.10).

Виджет Box Plot используется для наблюдения за статистическими свойствами набора данных. Также Box Plot полезен для поиска свойств определенного набора данных, например, набора экземпляров, вручную определенных в другом виджете (например, Scatter Plot или экземпляров, принадлежащих некоторому кластеру или узлу дерева классификации).

Рис.10. Добавление виджета Box Plot

11) Для того что визуализировать данные необходимо соединить виджет File и Box Plot (Рис.11).

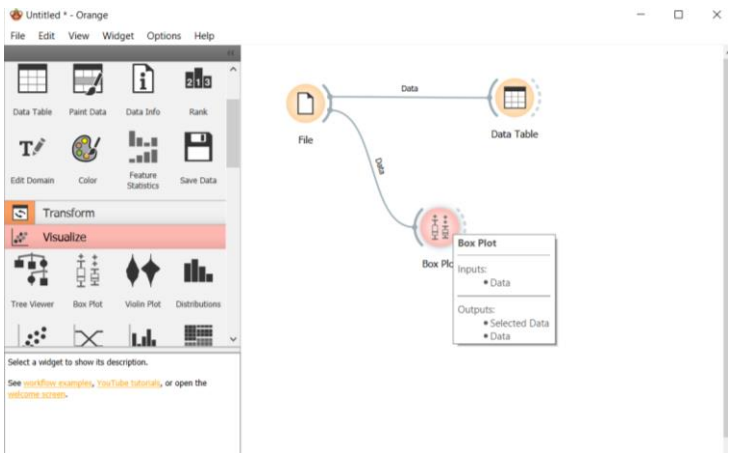


Рис.11. Соединение виджета File и Box Plot

12) Для просмотра визуализации данных необходимо открыть виджет Box Plot на холсте (Рис.12).

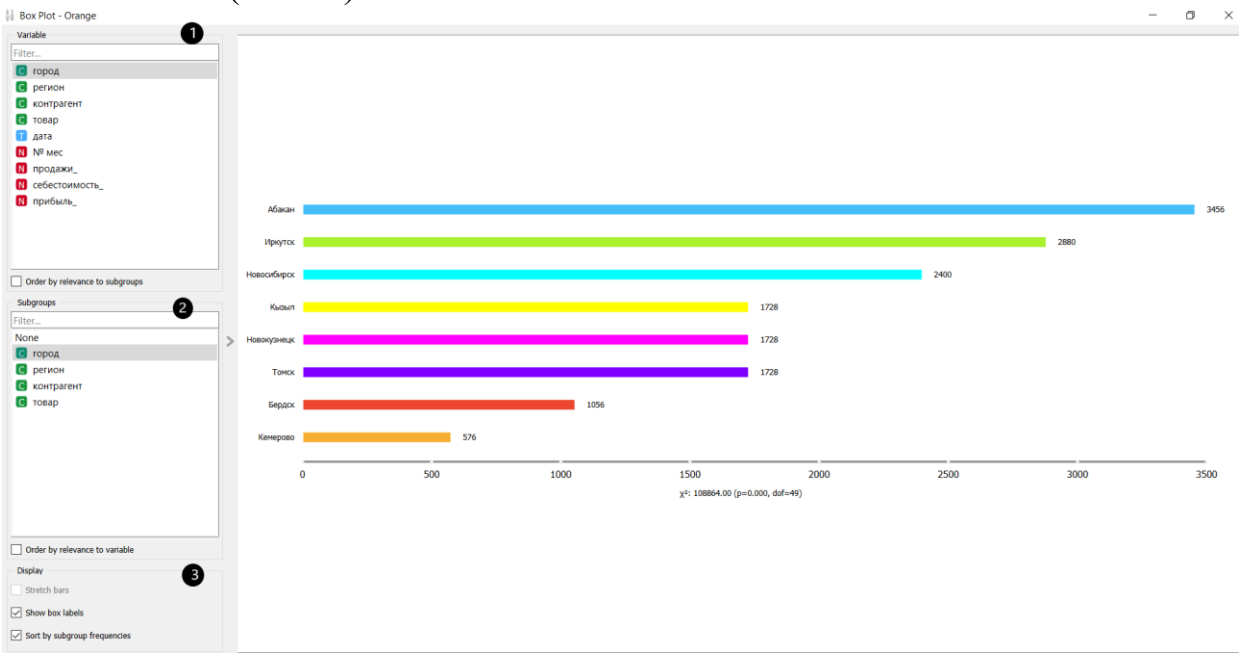


Рис.12. Диалоговое окно виджета Box Plot

Описание диалогового окна Box Plot (Рис.12).

1. Выбор переменной, которую нужно отобразить. Установите галочку порядок по релевантности подгруппам для переменных порядка по Chi2 или ANOVA над выбранной подгруппой.

2. Выбор Подгруппы, чтобы просмотреть графики полей, отображаемые дискретной подгруппой. Установите галочку порядок по релевантности переменной к подгруппам порядка по Chi2 или ANOVA над выбранной переменной.

3. Если экземпляры сгруппированы по подгруппе, можно изменить режим отображения. Аннотированные поля будут отображать конечные значения, среднее и медиану, в то время как сравнение медиан и сравнение средних, естественно, будет сравнивать выбранное значение между подгруппами.

13) С помощью Box Plot проанализируем данные о продажах по городам. Для этого в variable выбираем переменную продажи, а также в subgroups выберем подгруппу город. В итоге получили данные о продажах по городам(Рис.13)

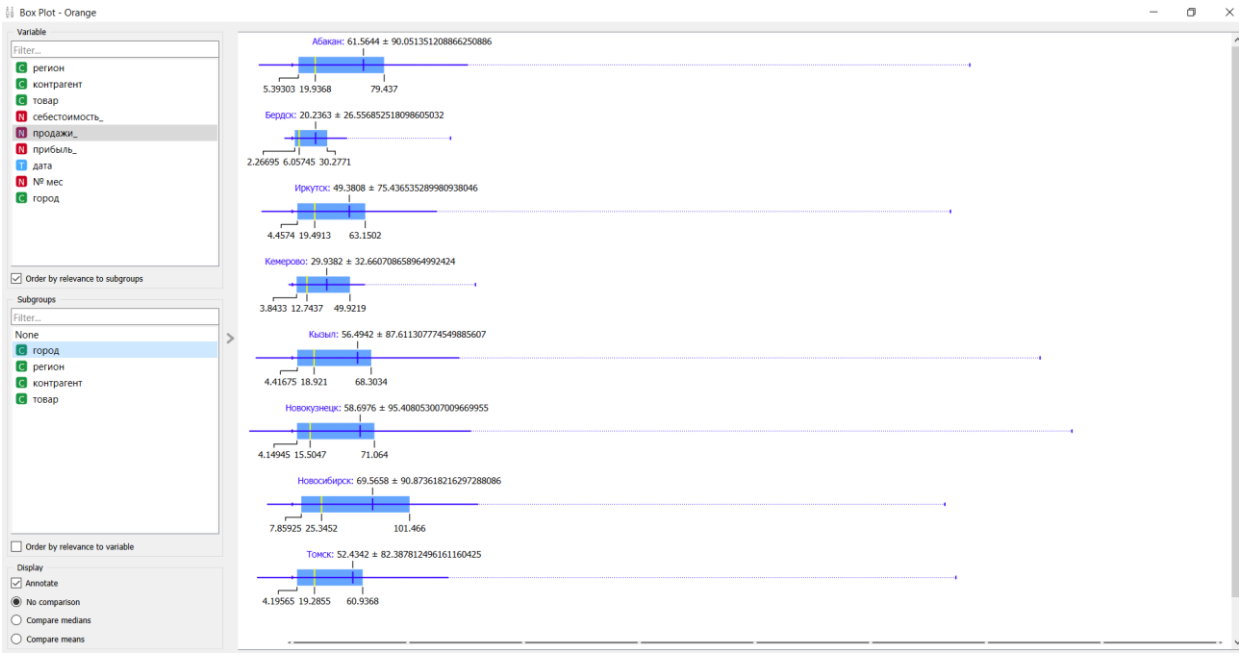


Рис.13. Данные о продажах по городам

14) Также для разведочного анализа визуализировать данные можно с помощью виджета Distributions (Рис.14).

Distributions- виджет отображающий распределение значений дискретных или непрерывных атрибутов. Если данные содержат переменную класса, распределения могут быть обусловлены классом.

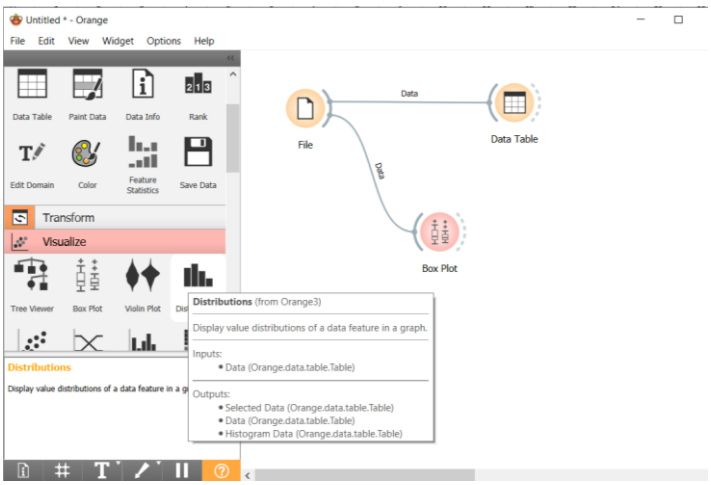


Рис.14. Виджет Distributions

15) Добавляем виджет Distributions на холст и открываем (Рис.15).

Рис.15. Добавление и открытие виджета Distributions

16) Открылось диалоговое окно виджета Distributions (Рис.16).

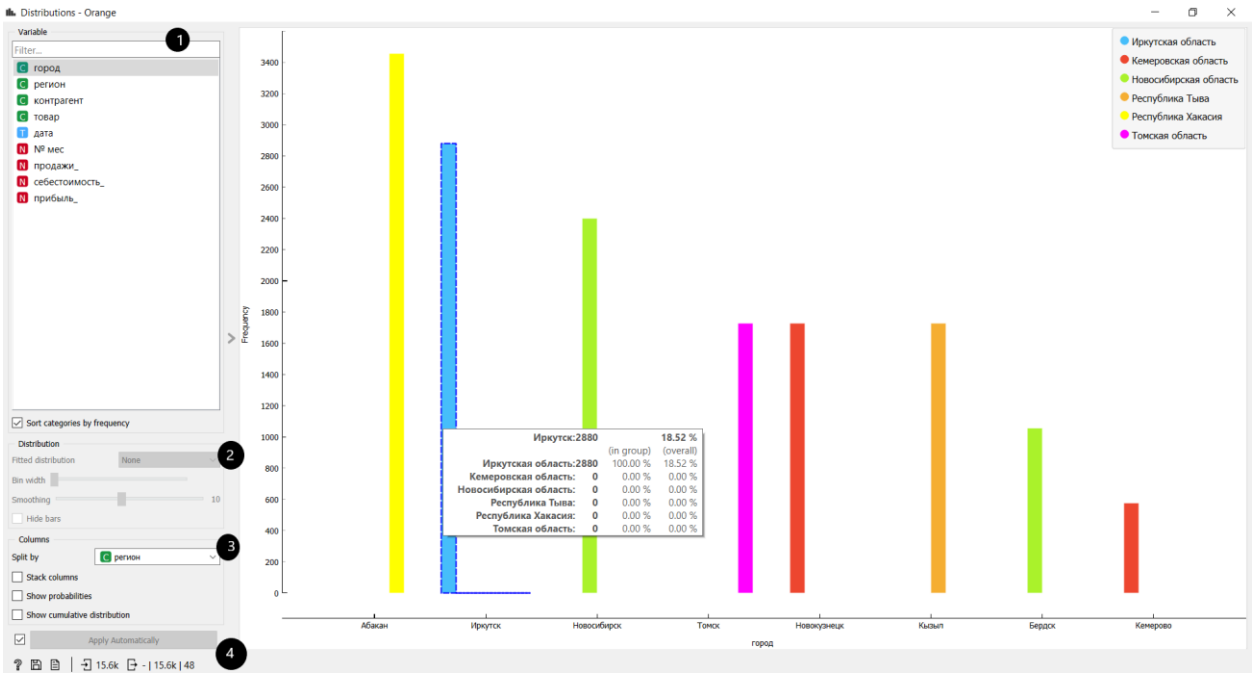


Рис.16. Диалоговое окно виджета Distributions

Описание диалогового окна Distributions (Рис.16).

1. Список переменных для отображения. Сортировка категорий по порядку частот, отображаемые значения по частоте.

2. Установка ширины корзины с помощью ползунка. Прецизионная шкала устанавливается на разумные интервалы. Встроенное распределение подходит выбранному распределению к графику. Варианты: Нормальная, Бета, Гамма, Рэлея, Парето, Экспоненциальная, Плотность ядра.

3. Столбцы:

- Разделение по отображает распределения значений для экземпляров определенного класса.

- В столбцах стека отображается один столбец на корзину, окрашенный пропорциями значений класса.

- Показать вероятности показывает вероятности значений классов в выбранной переменной.

- Отображение кумулятивного распределения кумулятивно складывает частоты.

4. Если установлен флажок применить автоматически, изменения сообщаются автоматически. Также можно нажать кнопку Применить.

17) С помощью виджета Distributions проведем анализ данных количества товаров контрагентов. Для этого необходимо выбрать в variable «товар», а в columns «контрагент». В итоге можно просмотреть общее количества товаров брендов, а также количество товаров контрагентов (Рис.17).

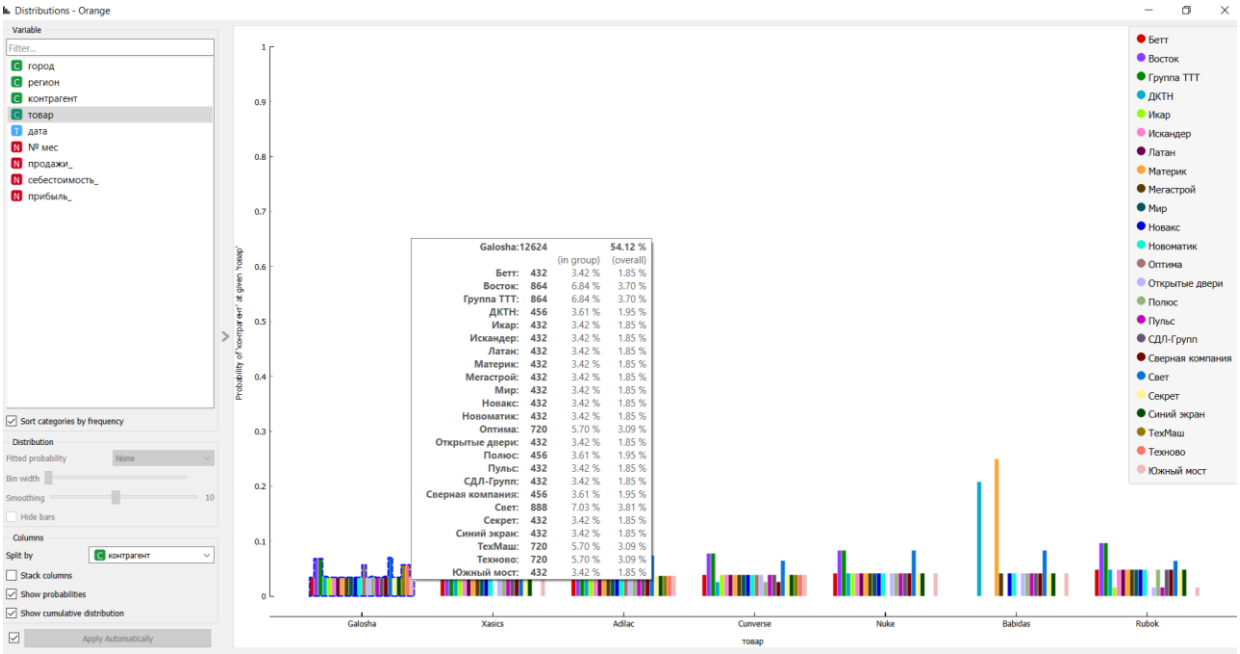


Рис.17. Данные о количестве товара контрагентов

18) Визуализировать данные можно и с помощью Scarlett plot (Рис.18).

Scarlett plot предоставляет 2-мерную визуализацию точечной диаграммы. Данные отображаются в виде набора точек, каждая из которых имеет значение атрибута оси X, определяющего положение на горизонтальной оси, и значение атрибута оси Y, определяющего положение на вертикальной оси. Различные свойства графика, такие как цвет, размер и форма точек, заголовки осей, максимальный размер точки и дрожание, могут быть отрегулированы в левой части виджета.

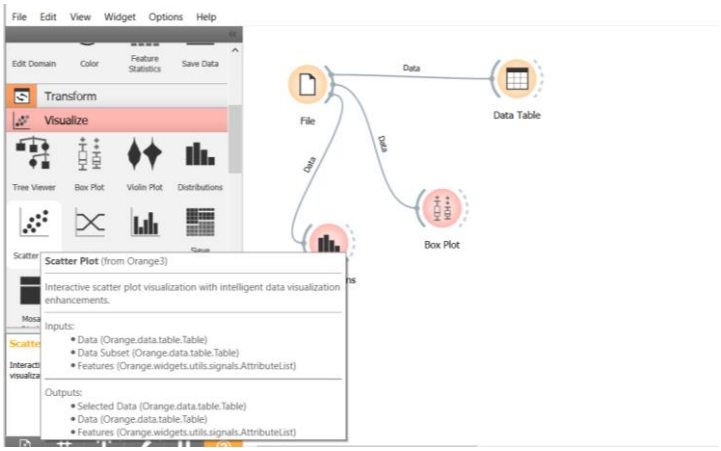


Рис.18. Виджет Scatter Plot

19) Добавляем Scatter Plot на холст и соединяем с виджетом File (Рис.19).

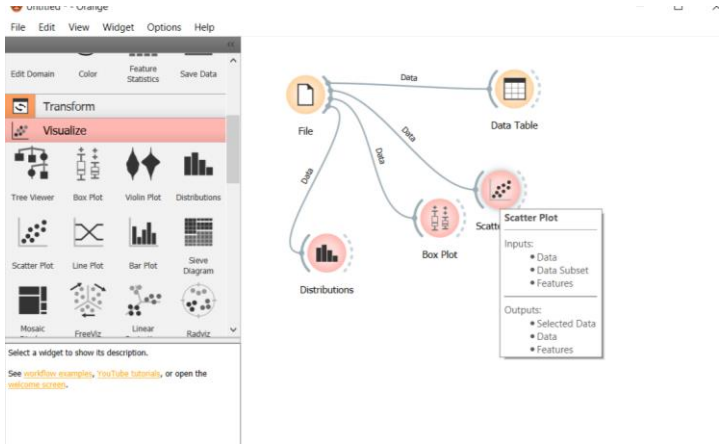


Рис.19. Добавление Scatter Plot и соединение с виджетом File

20) Открываем Scatter Plot (Рис.20).

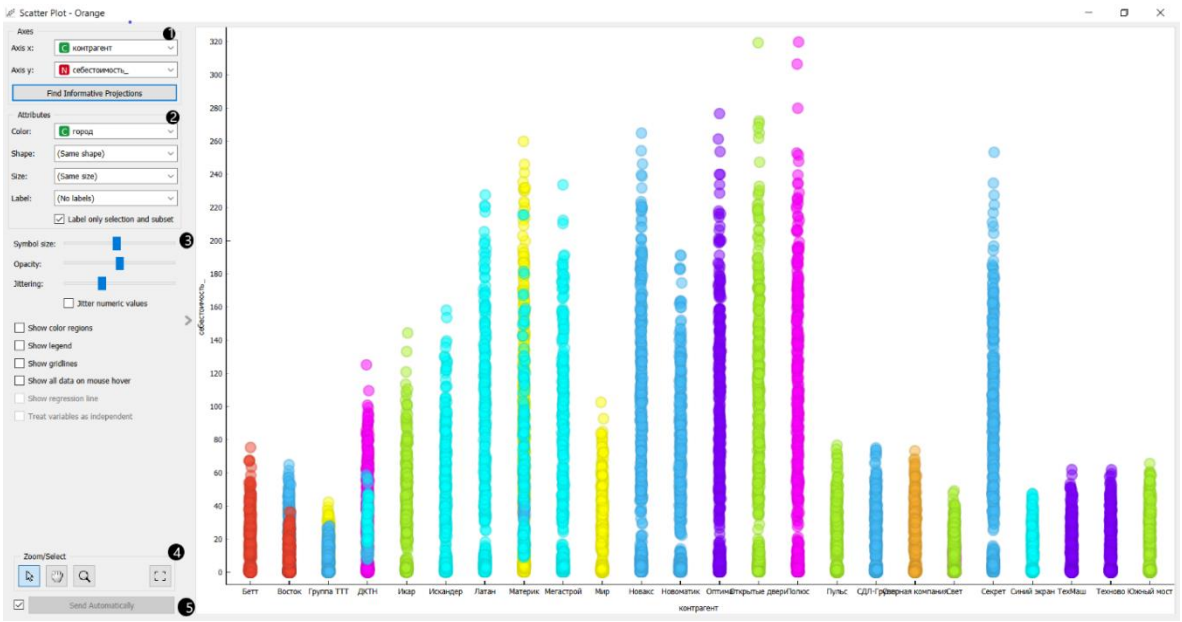


Рис.20. Диалоговое окно Scatter Plot

1. Выбор атрибутов x и y. Оптимизируйте свою проекцию с помощью функции поиска информативных проекций. Эта функция оценивает пары атрибутов по средней точности классификации и возвращает пару с наибольшим баллом с одновременным обновлением визуализации.

2. Атрибуты: Задайте цвет отображаемых точек (вы получите цвета для категориальных значений и сине-зелено-желтые точки для числовых). Задайте метку, форму и размер, чтобы различать точки. Метка только выбранных точек позволяет выбирать отдельные экземпляры данных и помечать только их.

3. Задайте размер и непрозрачность символа для всех точек данных. Установите дрожание, чтобы предотвратить перекрытие точек. Дрожание будет случайным образом рассеивать точки только вокруг категориальных значений. Если установлен флажок Числовые значения Джиттера, точки также рассеиваются вокруг их фактических числовых значений.

- Отображение цветовых областей цветов графика по классам.
- Показать легенду отображает легенду справа. Щелкните и перетащите легенду, чтобы переместить ее.
- Показать линии сетки отображает сетку за графиком.
- Показывать все данные при наведении указателя мыши позволяет создавать информационные пузырьки, если курсор помещен на точку.
- Показать линию регрессии рисует линию регрессии для пары числовых атрибутов. Если для раскраски графика выбрана категориальная переменная, будут отображаться отдельные линии регрессии для каждого значения класса.
- Отношение к переменным как к независимым соответствиям линии регрессии к группе точек (минимизация расстояния от точек), а не как к функции x (минимизация вертикальных расстояний).

4. Выбор, масштабирование, панорамирование и масштабирование по размеру - это варианты изучения графика. Ручной выбор экземпляров данных работает как инструмент углового/квадратного выделения. Дважды щелкните, чтобы переместить проекцию. Прокрутка вверх или вниз для масштабирования.

5. Если установлен флажок отправить автоматически, изменения сообщаются автоматически. Также можно нажать кнопку Отправить.

21) Проведем анализ данных продаж по месяцам используя Scarlett plot. Выберем атрибуты в axes x «продажи», а в axes y «№ мес». С помощью графика можно определить сколько было продаж в месяц (Рис.21).

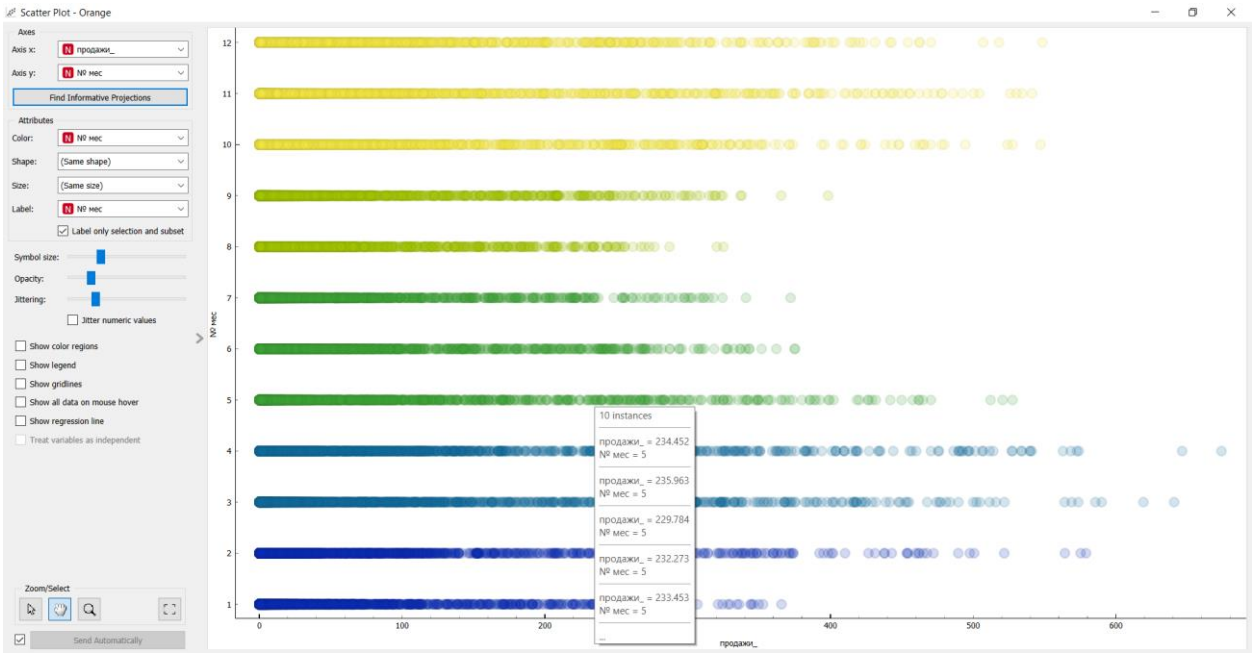


Рис.21. Количество продаж в месяц

22) В итоге получилась готовая схема разведочного анализа (Рис.22).

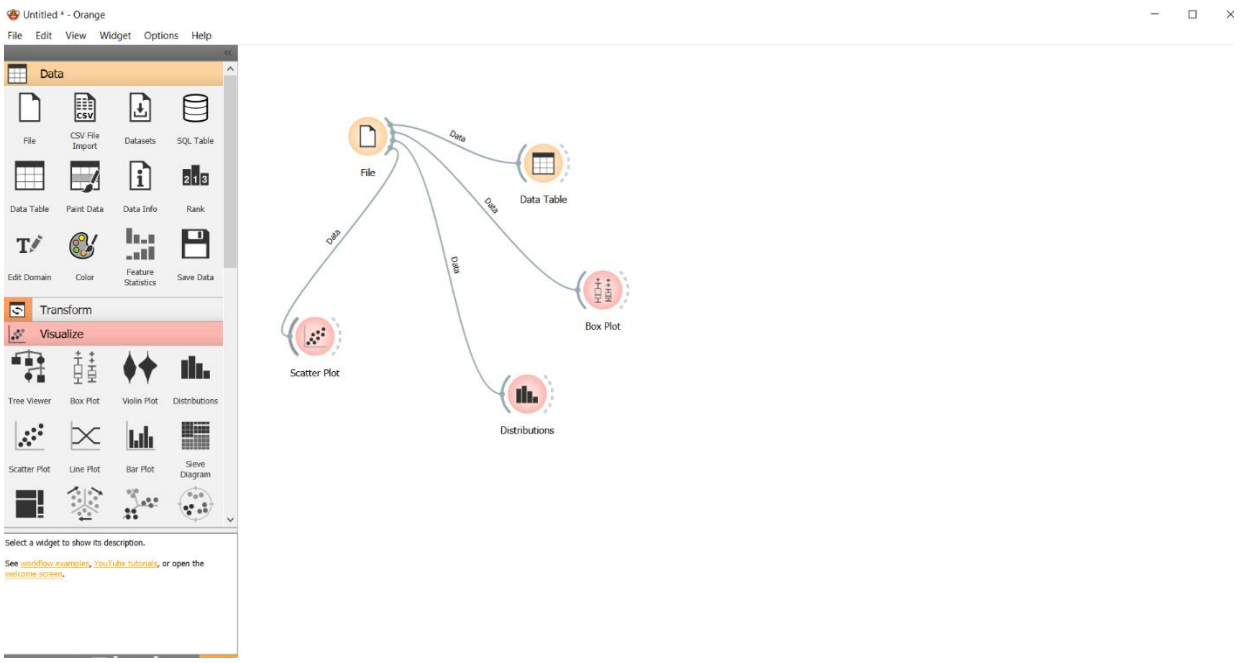


Рис.22. Итоговая схема

Выводы

В данной работе был выполнен разведочный анализ данных с помощью программного пакета визуального программирования на основе компонентов для визуализации данных Orange. С помощью виджетов File, Data Table, Distributions, Scatter Plot и Box Plot выполнили анализ данных о продажах по городам, о количестве товара контрагентов, количество продаж в месяц.

Библиографический список

1. Мастевной С. С., Петрова А. Н. Data mining: обзор методов и области их применения // Наука, инновации и технологии: от идей к внедрению. 2022. С. 38-40. URL: https://www.elibrary.ru/item.asp?id=48375089
2. Максимов В. Е., Резникова К. М., Попов Д. А. Информационные технологии для анализа данных морского флота // Отходы и ресурсы. 2021. Т. 8. №. 1. С. 6-6. URL: https://www.elibrary.ru/item.asp?id=45420956
3. Омарова Ш. Е., Медеубаева А. М. Сравнительный анализ инструментов data mining // Заметки ученого. 2020. №. 11. С. 185-193. URL: https://www.elibrary.ru/item.asp?id=44642833
4. Маматкасымова А.Т., Кульматова Н.А. Orange: Использование системы визуального программирования при обработке больших данных// Материаловедение. 2022. №1(36). С. 2232. URL: https://www.elibrary.ru/item.asp?id=49202512
5. Кривенко Ю. С., Минасян А. Т., Разиньков А. О. Исследование технологий интеллектуального анализа данных (Data Mining) // Актуальные проблемы управления в электронной экономике. 2018. С. 182-184. URL: https://www.elibrary.ru/item.asp?id=37071249
6. Пальмов С. В., Жуйкова А. А. Поиск ассоциативных правил средствами аналитического пакета orange // Форум молодых ученых. 2018. №. 8. С. 533-538. URL: https://www.elibrary.ru/item.asp?id=36425043
7. Токарев А. И., Брякин А. Н. Разведочный анализ данных и data mining // Перспективные направления развития отечественных информационных технологий. 2017. С. 201-203. URL: https://www.elibrary.ru/item.asp?id=34962226