Дипломная работа по профессии Инженер данных

К работе приложены следующие файлы:

1. diploma\_task\_description.docx – описание задания
2. diploma\_afanasyev.docx – текущий файл, описание решения
3. supermarket\_sales - Sheet1.csv – предоставленный исходный датасет
4. tables\_creation.sql – sql скрипты по созданию всех таблиц в БД Postrges
5. select\_scripts.sql – sql скрипты для проверок таблиц и промежуточных результатов
6. dim\_tables\_output.ktr – ETL трансформация для наполнения таблиц измерений
7. fact\_tables\_output.ktr – ETL трансформация для наполнения таблицы фактов
8. nds\_supermarket\_job.kjb – общий ETL
9. er\_diagram.png – NDS схема (звезда)
10. Supermarket\_sales\_dashboard.twb – файл дэшбоардов
11. main\_dashboard.png – изображение основного дэшбоарда
12. sales\_dashboard.png – изображение вспомогательного дэшбоарда
13. Обработайте и проанализируйте данные

Данные из предлагаемого ресурса были выгружены в локальный файл формата CSV. Для оценки и анализа данных был использован исходный сайт и предварительно загруженная таблица в базе данных Postgres в локальном контейнере Docker.

Описание данных приведено в исходном источнике. Для нас интересны следующие моменты:

* тип данных – это выгрузка по продажам разных товаров одной компании в трех разных городах, которые соответствуют трем разным филиалам, существует строгая привязка города к филиалу, что в свою очередь делает эти два измерения идентичными и одинаковыми;
* период выгрузки – указано, что продажи выгружены за три месяца, т.к. у нас отсутствует начало и окончание периода, мною было принято решение не использовать версионирование в таблице фактов;
* количество записей – указано точно, что предоставлено 1000 уникальных записей инвойсов, что дает нам использовать данное поле, как основной ключ в таблице фактов;
* среди 17-и столбцов можно легко проследить, что поля: branch, city, customer type, gender, product line и payment – это признаки и эти данные будут использоваться, как измерения;
* поля unit price, quantity, tax 5%, total date, time, cogs gross margin percentage, gross income, rating, invoice id и колонки с суррогатными ключами таблиц измерений будут использоваться для наполнения таблицы фактов.

В источнике приведены процентные соотношения по каждому измерению, из которых можно сделать вывод, что предоставленные данные очень консистентные и почти эквивалентны в процентном соотношение почти по всем типам фактов, что в итоге сказалось на дэшбоардах, в которых многие столбцы, по выбранным измерениям, имеют одинаковую величину.

1. Сформируйте нормализованную схему данных (NDS)

Как было упомянуто в прошлом блоке, следующие столбцы были выбраны для формирования таблиц:

* Измерения/витрины: branch, city, customer type, gender, product line и payment. Соответственно всем переменным был выбран тип данных: String (строковые данные)
* Факты: unit price, quantity, tax 5%, total date, time, cogs gross margin percentage, gross income, rating, invoice id. В таблице фактов для даты был выбран тип данных: date, а для quantity и суррогатных ключей – integer.
* Вспомогательная таблица для хранения описание ошибок: CSV reading errors – String.

Для всех измерений были наложены ограничения на уникальность. Для таблицы фактов также было задано условие не иметь пустых строк.

Ниже представлена схема звезда для организации таблиц и их связей (DDS):

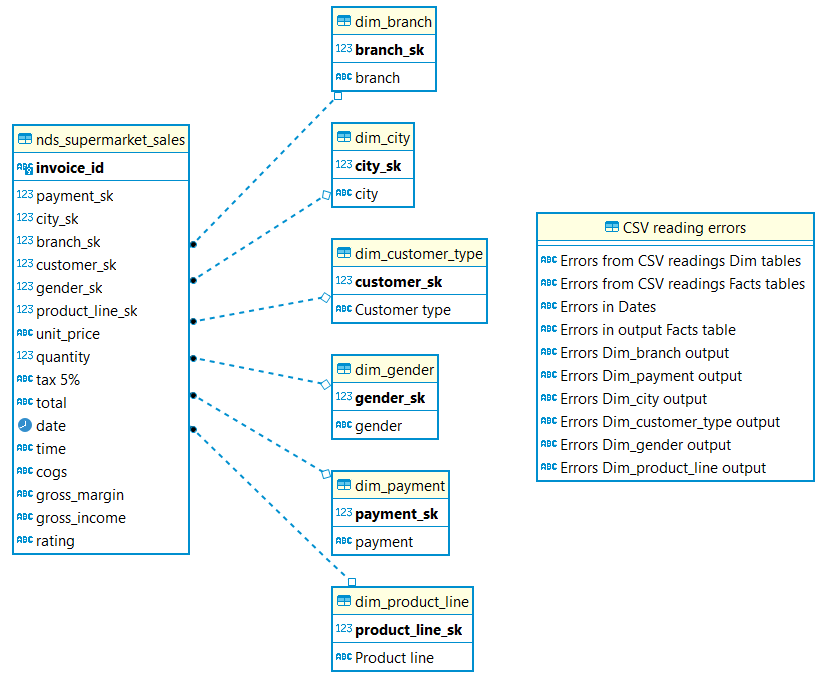


Рис.1 ER диаграмма

Таблица nds\_supermarket\_sales является нормализованной таблицей фактов. Для таблиц измерений / витрин были сгенерированы суррогатные ключи, которые были подставлены в таблицу фактов вместо самих измерений. Суррогатные ключи были сгенерированы на стадии создания таблиц с помощью sql запроса.

1. Сформируйте ETL-процессы: для заливки данных в NDS и для создания витрин

Для формирования ETL процесса были созданы две отдельные трансформации, которые были объединены в работу JOB в pentaho.

1. **Процесс преобразования и наполнения таблиц измерений / витрин:**

* Загрузка данных из локального / либо, хранящегося удаленно, CSV-файла;
* Сортировка данных по отдельному измерению для каждой ветке;
* Удаление дубликатов;
* Запись в таблицу измерений dim\_\*.
* Процесс распараллелен для 6-и измерений.

**Параметры процесса**

Параметры отсутствуют.

**Вход процесса**

Источником записей для процесса является CSV-файл supermarket\_sales - Sheet1.csv.

Перечень полей CSV-файл supermarket\_sales - Sheet1.csv:

|  |  |  |
| --- | --- | --- |
| Поле | Краткое описание | Тип данных |
| Invoice ID | Уникальный инвойс - Первичный ключ | STRING |
| Branch | Тип филиала | STRING |
| City | Город продажи | STRING |
| Customer type | Тип покупателя | STRING |
| Gender | Пол покупателя | STRING |
| Product line | Тип продуктовой линии | STRING |
| Unit price | Цена единицы товара | NUMBER |
| Quantity | Количество товаров, купленное покупателем | INTEGER |
| Tax 5% | Налог, неоплачиваемый покупателем | NUMBER |
| Total | Стоимость, включая налоги | NUMBER |
| Date | Дата покупки | DATE |
| Time | Время покупки | STRING |
| Payment | Тип оплаты покупки | STRING |
| Cogs | Стоимость корзины покупателя | NUMBER |
| Gross margin percentage | Размер налога | NUMBER |
| Gross income | Размер дохода | NUMBER |
| Rating | Рейтинг удовлетворенности покупателя | NUMBER |

**Алгоритм работы**

1. Выполняется чтение строк из CSV-файл supermarket\_sales - Sheet1.csv.
   1. На шаге настроен контроль качества -> если во время чтения файла произойдет ошибка, то текст ошибки будет направлен в отдельный поток с полем Errors from CSV readings Dim tables
   2. В потоке обработке ошибок произойдет обрезка пробелов вокруг Сообщения ошибки
   3. Далее произойдет замена переноса строки на символ решетки
   4. Будет произведена запись в специально созданную таблицу CSV reading errors
2. Сортировка потока по бизнес ключу в возрастающем порядке.
3. Выбор уникальных строк в потоке
4. Запись в таблицу измерения / витрину dim\_\*
   1. На шаге настроен контроль качества -> если во время записи файла произойдет ошибка, то текст ошибки будет направлен в отдельный поток с полем Errors Dim\_\* output
   2. В потоке обработке ошибок произойдет обрезка пробелов вокруг Сообщения ошибки
   3. Далее произойдет замена переноса строки на символ решетки
   4. Будет произведена запись в специально созданную таблицу CSV reading errors

Потоки трансформации для каждого измерения со своими названиями идентичны.

Выход процесса: Наполненные шесть таблиц измерений – витрин и одна таблица с сообщениями ошибок, если таковые произойдут на этапе трансформации.

Перечень полей таблицы dim\_branch:

|  |  |  |
| --- | --- | --- |
| Поле | Краткое описание | Тип данных |
| branch\_sk | Суррогатный ключ записи | INTEGER |
| branch | Тип филиала, уникальные значения | STRING |

Перечень полей таблицы dim\_city:

|  |  |  |
| --- | --- | --- |
| Поле | Краткое описание | Тип данных |
| city\_sk | Суррогатный ключ записи | INTEGER |
| city | Название города, уникальные значения | STRING |

Перечень полей таблицы dim\_customer\_type:

|  |  |  |
| --- | --- | --- |
| Поле | Краткое описание | Тип данных |
| customer\_sk | Суррогатный ключ записи | INTEGER |
| Customer type | Тип покупателя, уникальные значения | STRING |

Перечень полей таблицы dim\_gender:

|  |  |  |
| --- | --- | --- |
| Поле | Краткое описание | Тип данных |
| gender\_sk | Суррогатный ключ записи | INTEGER |
| gender | Пол покупателя, уникальные значения | STRING |

Перечень полей таблицы dim\_payment:

|  |  |  |
| --- | --- | --- |
| Поле | Краткое описание | Тип данных |
| payment\_sk | Суррогатный ключ записи | INTEGER |
| payment | Вид оплаты, уникальные значения | STRING |

Перечень полей таблицы dim\_product\_line:

|  |  |  |
| --- | --- | --- |
| Поле | Краткое описание | Тип данных |
| product\_line\_sk | Суррогатный ключ записи | INTEGER |
| Product line | Тип продуктовой линейки, уникальные значения | STRING |

1. **Процесс преобразования и наполнения таблицы фактов:**

* Загрузка данных из локального / либо, хранящегося удаленно, CSV-файла;
* Выбор полей и проверка даты;
* Подмена факта на суррогатный ключ по бизнес ключу;
* Повторение подмены на суррогатные ключи для каждого измерения (5 раз);
* Запись в таблицу фактов nds\_supermarkets\_sales.

**Параметры процесса**

Параметры отсутствуют.

**Вход процесса**

Источником записей для процесса является CSV-файл supermarket\_sales - Sheet1.csv, а также, ранее созданные таблицы измерений – витрины: dim\_branch, dim\_city, dim\_customer\_type, dim\_gender, dim\_payment, dim\_product\_line.

Перечень полей таблицы nds\_supermarket\_sales:

|  |  |  |
| --- | --- | --- |
| Поле | Краткое описание | Тип данных |
| Invoice ID | Уникальный инвойс - Первичный ключ | STRING |
| branch\_sk | Суррогатный ключ филиала | INTEGER |
| city\_sk | Суррогатный ключ для города | INTEGER |
| customer\_sk | Суррогатный ключ для типа покупателя | INTEGER |
| gender\_sk | Суррогатный ключ для пола | INTEGER |
| product\_line\_sk | Суррогатный ключ для типа продуктовой линии | INTEGER |
| Unit price | Цена единицы товара | NUMBER |
| Quantity | Количество товаров, купленное покупателем | INTEGER |
| Tax 5% | Налог, неоплачиваемый покупателем | NUMBER |
| Total | Стоимость, включая налоги | NUMBER |
| Date | Дата покупки | DATE |
| Time | Время покупки | STRING |
| payment\_sk | Тип оплаты покупки | INTEGER |
| Cogs | Стоимость корзины покупателя | NUMBER |
| Gross margin percentage | Размер налога | NUMBER |
| Gross income | Размер дохода | NUMBER |
| Rating | Рейтинг удовлетворенности покупателя | NUMBER |

**Алгоритм работы**

* 1. Выполняется чтение строк из CSV-файл supermarket\_sales - Sheet1.csv.

1.1 На шаге настроен контроль качества -> если во время чтения файла произойдет ошибка, то текст ошибки будет направлен в отдельный поток с полем Errors from CSV readings

1.2 В потоке обработке ошибок произойдет обрезка пробелов вокруг Сообщения ошибки

1.3 Далее произойдет замена переноса строки на символ решетки

1.4 Будет произведена запись в специально созданную таблицу CSV reading errors в поле Errors from CSV readings Facts Tables

2. Выбор полей в поток и конвертация даты для выполнения контроля поля Дат. Данный шаг можно убрать из трансформации, т.к. в шаге считывания файла, если произойдет ошибка она будет занесена в ошибку соответствующей таблицы.

2.1 На шаге настроен контроль качества -> если во время перевода дат произойдет ошибка, то текст ошибки будет направлен в отдельный поток с полем Errors in Dates

2.2 В потоке обработке ошибок произойдет обрезка пробелов вокруг Сообщения ошибки

2.3 Далее произойдет замена переноса строки на символ решетки

2.4 Будет произведена запись в специально созданную таблицу CSV reading errors в поле Errors in dates

3. Производится поиск и замена по бизнес ключу Branch суррогатного ключа в основном потоке из таблицы Dim\_branch. При этом основной столбец с бизнес ключом удаляется из потока.

1. Производится поиск и замена по бизнес ключу City суррогатного ключа в основном потоке из таблицы Dim\_city. При этом основной столбец с бизнес ключом удаляется из потока.
2. Производится поиск и замена по бизнес ключу Payment суррогатного ключа в основном потоке из таблицы Dim\_payment. При этом основной столбец с бизнес ключом удаляется из потока.
3. Производится поиск и замена по бизнес ключу Product line суррогатного ключа в основном потоке из таблицы Dim\_product\_line. При этом основной столбец с бизнес ключом удаляется из потока.
4. Производится поиск и замена по бизнес ключу Gender суррогатного ключа в основном потоке из таблицы Dim\_gender. При этом основной столбец с бизнес ключом удаляется из потока.
5. Производится поиск и замена по бизнес ключу Customer type суррогатного ключа в основном потоке из таблицы Dim\_customer\_type. При этом основной столбец с бизнес ключом удаляется из потока.
6. Далее производится запись в таблицу фактов. При этом если дата поля Date будет больше текущей даты, то произойдет ошибка записи и ошибка будет направлена в поток для обработки ошибок.
   1. На шаге настроен контроль качества -> если во время заливки данных в таблицу произойдет ошибка, то текст ошибки будет направлен в отдельный поток с полем Errors in output Facts table
   2. В потоке обработке ошибок произойдет обрезка пробелов вокруг Сообщения ошибки
   3. Далее произойдет замена переноса строки на символ решетки
   4. Будет произведена запись в специально созданную таблицу CSV reading errors в поле Errors in output Facts table
7. **Обе трансформации сведены в работу Job для выполнения последовательно.**

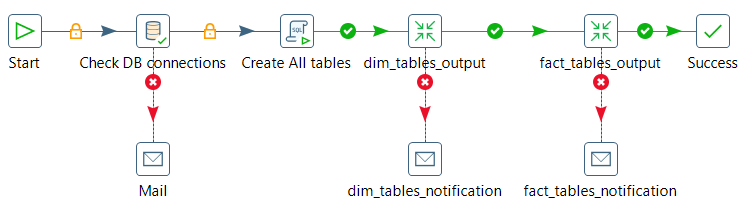


Рис. 2 Финальная последовательность для выполнения ETL.

Процесс выполнения финальной работы:

* Проверка подключения к базе данных
* Создание таблиц фактов и измерений
* Выполнение трансформации наполнения таблиц измерений / витрин
* Выполнение трансформации наполнения таблиц фактов

Алгоритм работы:

1. Работа Job начинается со стартового блока, в котором можно задать расписание для выполнения работы.
2. Шаг проверки подключения к базе данных Postgres.
   1. Если соединение отсутствует, происходит извещение администратора по электронной почте (в работе пароль для почты удален умышленно)
3. Далее происходит создание таблиц, соответствующих ER-диаграмме и вышеописанным таблицам
4. Выполнение трансформации по заливке в таблицы измерений, описанной в пункте А.
   1. Если произойдет ошибка выполнения трансформации, происходит извещение администратора по электронной почте (в работе пароль для почты удален умышленно)
5. Выполнение трансформации по заливке в таблицы фактов, описанной в пункте В.
   1. Если произойдет ошибка выполнения трансформации, происходит извещение администратора по электронной почте (в работе пароль для почты удален умышленно)
6. Завершение работы.

Контроль качества.

В etl процесс, а также в таблицы внедрены следующие пункты по контролю качества данных:

* Для таблиц измерений наложены ограничения на уникальность
* В таблице фактов заданы параметры отсутствия данных, в противном случае такие данные не попадут в итоговую таблицу, а в таблице с ошибками появится запись о проблеме
* На все шаги считывания файлов и записи в таблицы применены шаги отслеживания ошибок в шагах с последующем логированием в отдельной таблице
* Для поля Даты применена проверка, чтобы запись не могла быть больше текущей даты
* В работу Job внедрены шаги по проверке подключения к базе данных, а также проверки выполнения трансформаций и при возникновение ошибок на этих шагах должно уходить электронное сообщение администратору

**4. Сформируйте набор метрик и дашбордов на их основе**

На основе полученных данных были созданы два дэшбоарда. Один содержит все итоговые данных продаж, с возможностью применения фильтров по разным метрикам для анализа периода продаж Рис.3.

Основными элементами дэшбоарда являются:

* Общая сумма продаж
* Количество проданных единиц товаров
* Средний рейтинг удовлетворенности покупателей

В разбивках по категориями представлены следующие метрики:

* Разбивка по продажам в зависимости от типа оплаты
* Разбивка по продажам по типу покупателей
* Разбивка по продажам по полу
* Разбивка по продажам по продуктовой категории

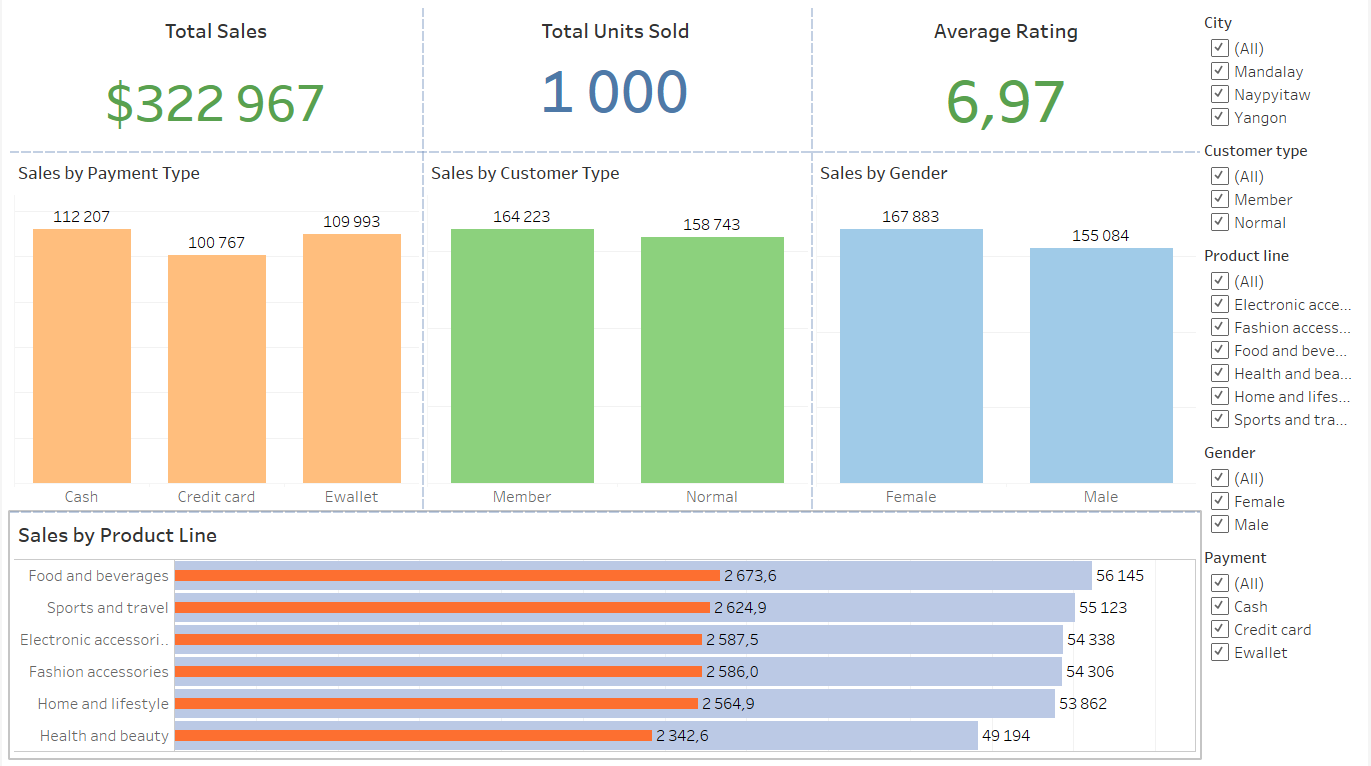


Рис.3 Основной дэшбоард

Во втором дэшбоарде приведена в основном информация о трендах продаж в разрезе времени и с возможностью применения фильтров Рис.4.

В дэшбоарде представлены два временных графика:

* Продажи по часам
* Продажи по месяцам

График по месяцам можно в будущем будет переделать на горизонтальный, т.к. с увеличением количества данных и месяцев график будет не удобным для чтения.

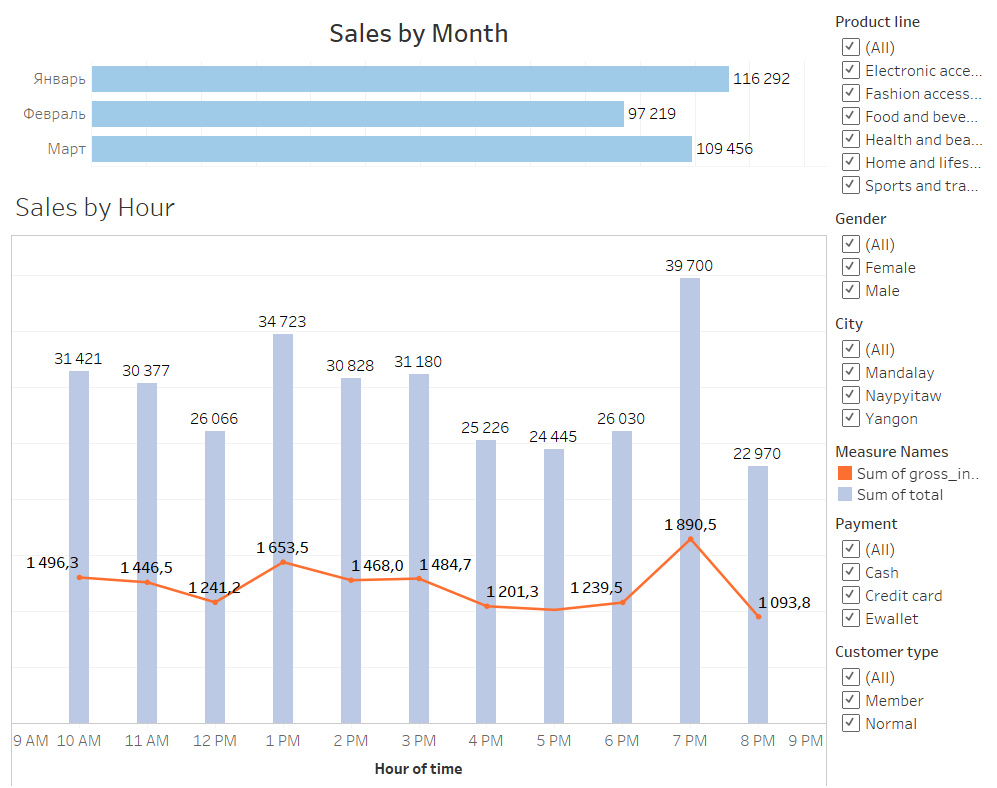


Рис. 4 Дэшбоард трендов продаж

5. Сформулируйте выводы

5.1 Выводы по ETL трансформации

Исходя из задания и представленных данных, некоторые процессы невозможно было реализовать. Такие, как создание версионирования в финальной таблице фактов. Для реализации такого процесса необходимо добавить в исходные данные о начале и окончание периода выгрузки данных.

Также в задание присутствовало ограничение в виде пункта создания таблиц, что в целом не предусматривало обновление таблиц измерений, либо не было сказано явно.

5.2 Выводы из анализа данных и графиков полученных в дэшбоардах

Представленные данные очень пропорционально распределены по каждому измерению. Диспропорция во многих измерениях наблюдается в 1-2%.

Довольно большое различие наблюдается в продажах продуктовых линеек. Хуже продаются товары для здоровья и красоты. Также можно сделать вывод, что покупатели предпочитаю расплачиваться наличными деньгами и электронными кошельками больше, чем кредитными картами.

Временной анализ нам дает вывод, что больше всего продаж совершается в 7 часов вечера и в обеденное время. По представленным месяцам сложно увидеть какую-то разницу в объемах продаж. Февральский спад обусловлен коротким месяцем.