1. **Какие проблемы в данных вы обнаружили и устранили перед анализом?**

|  |  |
| --- | --- |
| Проблема | Решение |
| Дубликаты в исходных файлах | Исключил дубликаты на стадии обработки в power query |
| Описание артикула – большое текстовое поле, много раз повторяется | Вынес в отдельный справочник items\_description, связал с основным справочником |
| В исходной таблице item информация о свойствах товара сохранена в формате коллекции (json или словарь). Обработка данной структуры достаточно сложна в power query, а при обработке требуется привести эти данные в вид плоской таблицы для возможности анализа по свойствам (если потребуется). | Я предположил, что по некоторым значениям атрибутов пользователю потребуется делать аналитику (например страна производитель), поэтому эти данные требуется обработать.  Написал в python функцию для распаковки коллекции (Task1\_unpack\_collection.py), т.к. это конструкция похожа на словарь.  Запустил скрипт в Power Query, дальше работал уже с более удобным датасетом.  Это большой объем данных, потому вынес значения категорий в отдельный справочник - items\_category, со столбцами Артикул, Категория, Значение категории.  Замечание. Все категории кроме 4 содержат по 1 атрибуту, кроме  NAZNACHENIE  UHOD\_ZA\_TKANJU  SUBTYPE  UZOR  Тут сложно предположить без общения с конечным пользователем, какую информацию он захочет получать из этих значений.  Менять с латиницы на русский язык не стал, если требуется, это реализуется быстро. |
| Одна из категорий начинается не с латинской буквы а с символа. \_differentsialnoe\_peremeschenie | В Power Query заменил на differentsialnoe\_peremeschenie |
| В справочнике item есть артикулы, которым не назначены рубрики в таблице menu | Описал при решении пункта 4 |
| В справочнике menu есть артикулы, у которых нет подрубрик или подразделов.  Я понимаю этот раздел как навигацию по сайту (глубину навигации) | Сделал дополнительные вычисляемые поля в справочнике menu – навигация (вида  сетка >>> фатин >>> мягкий ) и глубину навигации.  Возможно, это может пригодиться для дальнейшего анализа |
| У некоторых item отсутствуют основные цвета (всего 21 позиция) |  |

1. Какие группы товаров наиболее и наименее популярны? По каким параметрам можем объединить товары в группы для анализа? Создайте визуализацию для анализа закупщиком.

Группы из справочника items\_category, которые не стал использовать, т.к. для одного товара есть несколько таких категорий:

* Назначение
* Уход за тканью
* Подтип (subtype)
* Узор

Либо эти группы надо анализировать отдельно, но требуется ТЗ от пользователя.

Группы, которые заполнены у большинства товаров и могут быть использованы для группировки:

* Основной состав
* Производитель
* Процент основного состава
* Секция
* Ширина
* Тип
* Плотность (два параметра)
* Нарезка

Менее заполнены, и тоже могут быть использованы:

* Доп.состав
* В рулоне
* Система скидок
* Old\_sostav

Плюс из основного справочника item показатель – цвет. Также при желании можно использовать полученное значение из menu – навигация и глубина навигации.

Допустим, буду анализировать по таким группам:

* Цвет ткани
* Тип ткани
* Основной состав
* Процент основного состава
* Плотность
* Ширина
* Нарезка

При этом, для оценки наиболее и наименее популярных будем использовать 15 наиболее и наименее популярных цветов, а остальные признаки – как визуализацию распределения по этим параметрам.

Итого собраны две страницы – для самых популярных и самых не популярных.

Альтернатива – можно собрать сводную таблицу по группам цвет-параметры (пример на вкладке Задача 2 – альтернатива).

1. **Какие слова, не относящиеся к названию рубрики чаще встречаются в описании “успешных” тканей, а какие наоборот у аутсайдеров?**

Эту задачу целиком решал в python.

Категорию “успешности” определял как перечень товаров в разрезе id, которые дают 80% всего объема продаж (категория A). Категорию неуспешности определял как категорию С в терминологии ABC-анализа (нижние 5% товаров).

Алгоритм работы:

* Создал список рубрик
* Объединил все описания
* Создал список предлогов, чтобы потом исключить из поиска
* Разделил все описания по словам
* Исключил предлоги
* Подсчитал частоту встречаемости слов, не входящих в список слов из списка рубрик
* Сохранил топ-30 результатов в файл

Аналогичный алгоритм для аутсайдеров.

Скрипты:

Task3\_1.py

Task3\_2.py

Результаты в файлах:

результаты\_анализа\_3\_1.CSV

результаты\_анализа\_3\_2.CSV

Возможность анализа слов, с участием однокоренных выражений посмотрел, но сразу отложил. Для этого раньше использовалась библиотека pymorphy2, но ее перестали поддерживать в python.3.11, а с новой pymorphy3 за короткий срок разобраться не удалось.

Возможность решить тут есть, но потребуется побольше времени.

1. **Создайте визуализацию для руководителя по закупкам, которая показывает дефициты и профициты товара на остатке в разрезе конкретного товара и рубрик, если известно, что плечо поставки товара на склад соответствует периоду n. Желательно, чтобы визуализация давала не только общую информацию, но и давала возможность принять конкретные решения по товарам.**

Таблица для принятия решения с полями:

|  |  |
| --- | --- |
| Колонка таблицы | Описание |
| Артикул | Текущий артикул из справочника |
| Остаток на складе | Текущий остаток по артикулу на складе |
| Продажи за период | Продажи по артикулу за период n, принятые за усредненные данные, на которые ориентируемся (потребность на следующий период n) |
| Дефицит или профицит | Разница между складским остатком и потенциальной потребностью. Отрицательное значение – дефицит, положительное профицит. |
| Рекомендация по товару | Случай 1. Потребность больше текущего остатка – требуется закупка, в рекомендации предложено закупить с учетом того, что принятые на момент закупки заказы потребуется обеспечить  Случай 2. Потребность меньше текущего остатка, но меньше, чем потребность на ближайшие 2n периода – следует запланировать закупку в ближайший n период  Случай 3. На складе запас больше,чем на ближайшие 2n периода – ничего делать не требуется  Можно предусмотреть комментарий для существенного профицита (например запас больше чем на 6n периодов) |

Навигация по разделу – фильтры по разделам сайта Рубрика, Подрубрика, Подраздел рубрики.

Важно! Существуют товары, не включенные в рубрики – вынесены на лист Задача 4 исключение.

1. Какие аналитические инсайды возникли при анализе данных, какие рекомендации можно дать закупщику, какие гипотезы имеет смысл проверить?

Какие мысли возникли:

* Проверить зависимость продаж от глубины навигации на сайте, может провести A/B тест (подвигать категории)
* Зависимость продаж от наличия видео

Прочее:

Есть числовые категории, значения в которых указаны текстом. Это может быть не очень удобно для работы (или это просто текстовая выгрузка):

* Нарезка
* В рулоне
* Отпускается

Возможно корректнее хранить отдельно числовые значения (допустим со справочником допустимых значений), и категорийные.

Не очень нравится, что на один артикул есть много назначений – удобно ли потом анализировать, куда покупатель использовал, или может быть нет такой задачи? Нет ли дубликатов – одежда для персонала и спец.одежда это не одно и то же?

Формат предоставления

Допускается использование любой BI-системы, любой системы работы с данными от

Power BI до Jupiter Notebook. Результат можно отправить на почту timofeev@star-tex.ru

в течение 5 рабочих дней, после получения тестового задания.

Срок рассмотрения тестового задания - 3 рабочих дня.