**Итоговое задание по курсу**

**“Метрики, гипотезы, точки роста”**

Используем базу транзакций интернет-магазина из Великобритании (доступна по [ссылке](https://docs.google.com/spreadsheets/d/1L1cOzBw-Muaf4YZnsqEOrw4kjEL5avx7SjcE6uzCt0Q/edit?usp=sharing). При исследовании данных можно использовать любые инструменты (например, Excel, Google-таблицы, Python и так далее).

Ознакомиться с данными и составить аналитический отчет.

1. **Описание данных (что отражают, качество, полнота, какие некорректности и аномалии)**

База транзакций представлена в виде таблицы. Анализ данных проводился в Excel.

Общее количество транзакций - 54 214 шт. Каждая строка представляет собой позицию (StockNo) товара (количество и цена за единицу товара) в конкретном заказе (invoiceNo) конкретного пользователя (customerID) в определенное время (InvoiceDate). Группируя по **InvoiceNo**,можно видеть полный заказ конкретного пользователя в определенное время. Группируя по **CustomerID** - все заказы конкретного пользователя за весь период, охватываемый рассматриваемой базой транзакций. База транзакций имеет географию - каждая позиция имеет атрибут **Country** (связан с местом заказа, так как есть шесть пользователей, заказывающие больше чем из одной страны).

Таблица содержит следующие данные:

* InvoiceNo - номер счета

шестизначное число формата 5\*\*\*\*\*  (за исключением 960 строк - 1,8% от общего количества строк -, где к номеру добавлен префикс “С”, формат - С5\*\*\*\*\*). Все номера счета с префиксом имеют отрицательный атрибут Quantity, но не все отрицательные Quantity имеют префикс в номере счета!

* StockCode - код товара. Преимущественно пятизначное число (48569), лишь в 5344 случаях добавлен суффикс в виде букв латинского алфавита, также в 301 случае имеет одно из следующих значений:

|  |  |
| --- | --- |
| **StockCode** | **Description** |
| M | Manual |
| POST | POSTAGE |
| DOT | DOTCOM POSTAGE |
| D | Discount |
| DCGS0070 | CAMOUFLAGE DOG COLLAR |
| C2 | CARRIAGE |
| gift\_0001\_30 | Dotcomgiftshop Gift Voucher £30.00 |
| S | SAMPLES |
| DCGSSGIRL | GIRLS PARTY BAG |
| BANK CHARGES | Bank Charges |
| DCGS0004 | HAYNES CAMPER SHOULDER BAG |
| AMAZONFEE | AMAZON FEE |
| gift\_0001\_10 |  |
| CRUK | CRUK Commission |

* Description - описание товара, у 134 позиций отсутствует описание, две позиции со значением “??”. Все 136 позиций содержатся в множестве с незаполненными CustomerID. В базе за отчетный период фигурирует 3393 уникальных описания товара.
* Quantity – количество единиц товара в заказе. Присутствуют отрицательные значения, по всей видимости – возвраты товара и сторно. Баланс товаров не сходится в рамках представленного временного периода, т.е. есть возврат отдельных товаров, которые согласно базе транзакций не были куплены. Количество строк с отрицательным quantity - 1083, уникальных по invoiceNo – 869
* InvoiceDate - дата заказа в формате MM/D/YY HH:MM. Представленный в базе период транзакций – с 01.12.2019 по 09.12.2020
* UnitPrice - цена за единицу товара. Неотрицательные числа (float). Нулевая стоимость у 225 позиций (строк)
* CustomerID – id покупателя. Не заполнено у 13571 позиции (25.03% от общего числа строк), пропуски
* Country – страна. У 36 позиций (менее 1%) установлен маркер “Unspecified”

1. **Описание примененного подхода к анализу данных (ограничения накладываемые представленными данными и соответствующие интерпретации, допущения)**

После ознакомления с данными видим, что данные не нормализованные. Нет единой системы оформления типовых процессов, в системе учета нет или не в полной мере настроены проверки ввода данных при оформлении заказов.

Все пропуски и опечатки составляют не более 2% от общего числа транзакций. Для анализа в отчете предлагается пренебречь всеми незаполненными полями и исключить их из рассмотрения. А именно: пропуски в **Description** и значения “Unspecified” в **Country.**

Исключение составляют пропущенные id пользователя. Доля пропусков составляет порядка 25%, при этом остальные данные как правило в таких строках заполнены. Можно использовать в отчете, но не в анализе поведения и рейтингах пользователей.

Для того чтобы укрупненно посмотреть географию заказов, в таблицу был добавлен столбец с укрупненным географическим положением.

Chart, histogram

Description automatically generated

Как видно, что несоизмеримая доля заказов приходится на Великобританию и Ирландию (сгруппированы в GB&I, также условно в эту группу отнесены Нормандские острова

Graphical user interface, text, application

Description automatically generated

Второй зоной по количеству заказов является Европа (страны EU сгруппированы в Europe).

Для выявления тенденций в том числе визуально, можно GB&I и Europe рассмотреть отдельно.

Например,

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

**Chart, line chart

Description automatically generated**

Можно визуально выделить увеличение количества заказов перед рождественскими и новогодними праздниками в GB&I и Europe. В остальных странах такая тенденция не просматривается.

Для всех зон кроме Великобритании и Европы данные несоизмеримо меньше, возможно это нужно отнести к болям Бизнес-заказчика: кто же не хочет завоевать американский рынок!?

Также в 2020 году в мире была активная фаза развития COVID-19, поэтому, возможно тенденции в некоторых процессах могли нарушить обычное направление.

1. **Описание бизнес-заказчика, его гипотетических болей и целей развития бизнеса.**

Бизнес-заказчик –– собственник интернет-магазина.

По объемам продаж компания классифицируется как маленькая.

По количеству уникальных позиций и объемам продаж по географическим зонам, кажется как будто нужна диверсификация в части каталога и рынков сбыта.

В наиболее крупных зонах Великобритании и Европе активность продаж в середине недели «проваливается».

Chart, bar chart

Description automatically generated

И так как цель любого бизнеса – максимизация прибыли! (достижение баланса интересов стейкхолдеров). Для всех зон кроме Великобритании и Европы данные несоизмеримо меньше, возможно это нужно отнести к болям Бизнес-заказчика: кто же не хочет завоевать американский рынок!?

1. **Описание стейкхолдеров отчёта с обоснованием выбранных метрик (системы метрик)**

Для развития рекомендуется масштабировать маркетинговое направление.

Так как онлайн-магазин уже хоть и незначительно представлен в других регионах кроме GB и EU, то вероятно нужно инвестировать в развитие новых географических направлений. Провести маркетинговое исследование, увеличить охват, вовлечение.

Провести рекламные маркетинговые программы по увеличению FR, CTR.

Почти 25% customerID содержат пропуски. Делать выводы о количестве повторных покупок считаю нецелесообразно, высокая погрешность.

Добавить в отчет LTV.

Так как список товаров достаточно узкоспециализированный, возможно, нужно диверсифицировать каталог, расширив торговые сегменты до смежных.

В связи с вышесказанным предположением сотрудники от маркетинга будут одними из основных стейкхолдеров обновленного отчета, а значит необходимо дополнить маркетинговыми метриками.

1. **Рекомендации по модели данных бизнес-заказчика (что вы считаете важным отслеживать в дополнение к тому, что уже имеется в датасете). Выводы и рекомендации для бизнес-заказчика.**

* Для того чтобы максимально эффективно отслеживать данные, нужно диагностировать работу сбора данных: наладить их качественный сбор, обработку и хранение.
* Так как компания пока относится к маленьким, следующий уровень развития – средняя, необходимо определить и унифицировать систему метрик в компании.
* При оптимизации хранения и обработки данных необходимо добавить проверки и ограничения, чтобы значимые показатели не могли быть не заполнены.(customerID не заполнен на 25%).
* Диверсифицировать каталог, сделать упор на развитие отстающих регионов (с учетом географической специфики).
* Посмотреть в сторону уменьшения процента возврата. Текущее значение – 5.6% за отчетный период.