**Краткое описание набора данных**

Набор данных Brazilian E-Commerce Public Dataset by Olist представляет собой данные о заказах, сделанных в маркетплейсе Olist Store (olist.com , Бразилия). Это простая случайная выборка всех заказов, получивших отзывы покупателей. В наборе данных содержится информация почти о 100 тыс. заказов с 2016 г. по 2018 г., сделанных на нескольких торговых площадках. У Olist нет собственного магазина, он представлен на торговых площадках партнеров: Amazon, Americanas, Carrefour, Casas Bahia, Extra, MadeiraMadeira, Mercado Livre, Ponto Frio, Shoptime, Submarino, Zoom.

Данные содержатся в 9 таблицах и позволяют узнать информацию от статуса заказа, стоимости товаров, стоимости доставки (фрахта) до местоположения клиента, общей стоимости заказа, атрибутов продукта и, наконец, отзывов, написанных клиентами. Кроме того, доступны данные о местоположении покупателей и продавцов (почтовые индексы, координаты широты и долготы).

Olist предоставляет возможность продавцам со всей Бразилии продавать свои товары через маркетплейс и отправлять их напрямую клиентам, используя партнеров Olist по логистике.

После оформления покупателем заказа в Olist Store продавец получает уведомление о необходимости выполнения этого заказа. После получения покупателем заказа (или наступлении предполагаемой даты доставки) покупатель получает по электронной почте опрос об удовлетворенности, в котором он может оценить покупку по шкале от 1 до 5 и оставить свои комментарии.

Существует еще один набор данных от Olist Store – Marketing Funnel by Olist – это результаты маркетинговой воронки от продавцов, которые заполняли запросы на контакт для продажи своих продуктов в Olist Store.

**EDA: разведочный анализ данных**

Файл olist\_store\_EDA. ipynb

ER-диаграмма набора данных доступна по ссылке: https://lucid.app/lucidchart/invitations/accept/d54dc06e-caea-45ce-8141-eb0772b87077

Данные по актуальным заказам (не отмененным покупателями и товарам в наличии) охватывают период с сентября 2016 г. по сентябрь 2018 г. При этом недоступны данные за весь ноябрь 2016 г. (запуск новой версии платформы), а данные за декабрь 2016 г. и сентябрь 2018 г. представлены одним заказом. Поэтому, фактически, для большинства задач можно ограничиться периодом с января 2017 г. по август 2018 г. включительно.

**Пропуски данных.**

В данных есть пропуски и дубли записей:

| № п/п | Таблица | Наличие пропусков | Наличие дублей | Комментарии |
| --- | --- | --- | --- | --- |
| 1 | customers | нет | нет |  |
| 2 | order\_items | нет | нет |  |
| 3 | order\_payments | нет | нет | в 3-ёх случаях способ оплаты не определен |
| 4 | order\_reviews | есть, 58% записей | нет | отсутствует более половины отзывов и большая часть заголовков |
| 5 | orders | есть, 3% записей | нет | пропуски в полях: order\_approved\_at - дата и время подтверждения платежа order\_delivered\_carrier\_date - дата и время передачи заказа в доставку order\_delivered\_customer\_date - фактические дата и время доставки заказа клиенту |
| 6 | products | есть, 2% записей | нет | в основном, пропуски в полях: product\_category\_name - название категории товаров (на португальском) product\_name\_lenght - кол-во символов в названии товара product\_description\_lenght - кол-во символов, извлеченных из описания товара product\_photos\_qty - количество фотографий товара |
| 7 | category | нет | нет | список категорий дополнен 2-мя записями англоязычных названий категорий |
| 8 | sellers | нет | нет |  |
| 9 | geo | нет | есть, 26% записей |  |

Для решения задачи сегментации покупателей эти пропуски, на мой взгляд, несущественны.

**Параметры распределения переменных.**

Из всех количественных переменных наибольший интерес представляют следующие показатели:

price - цена товара

freight\_value - стоимость доставки товара (если заказ содержит более одного товара, то разделяется между товарами)

payment\_value - стоимость транзакции (сделки)

Распределение этих переменных имеет ассиметричный характер с положительным эксцессом и правосторонней асимметрией (среднее значение превышает медиану). Скорее всего, это логнормальное распределение.

Частотные графики и коробчатые диаграммы переменных наглядно это демонстрируют.

Для указанных переменных существуют значения, которые можно отнести к числу потенциальных выбросов и аномальных значений. В тоже время четких критериев для отнесения к этому разряду значений не существует.

Возможные подходы к выявлению потенциальных выбросов представлены в файле olist\_store\_outliers.ipynb

Для указанных переменных (с учетом характера их распределений) рекомендуется использовать показатель среднего абсолютного отклонения от медианы (среднего абсолютного отклонения (MAD)).

**Корреляция данных**

Для изучения возможных корреляций в данных построена матрица корреляций.

Из всех обнаруженных зависимостей можно выделить следующие:

- обратная зависимость между оценкой удовлетворенности (review\_score) и фактическими сроками доставки (duration\_delivery), коэффициент корреляции: -0.3 (это производная переменная)

- прямая зависимость между стоимостью доставки (freight\_value) и весом заказа (product\_weight\_g), 0.61

**Основные результаты EDA**

. Услугами магазина за период воспользовались 96,1 тыс. покупателей

. Покупателями магазина являются жители всех субъектов страны (26 штатов и одного фед. округа)

- 90% покупателей приходится на 10 субъектов страны

. Больше всего покупателей из штата Сан-Паулу и одноименного города этого штата:

- из штата Сан-Паулу - 42% покупателей

- из города Сан-Паулу - 16% покупателей

. Большинство продавцов – также из штата Сан-Паулу и одноименного города этого штата:

- из штата Сан-Паулу - 60% продавцов

- из города Сан-Паулу - 22% продавцов

. Количество категорий товаров: 73

. Наиболее представительная по кол-ву товаров категория: мебель

. Подавляющее большинство покупателей - 97% - совершило только одну покупку за период

. Самый популярный по кол-ву заказов способ оплаты: кредитная карта

. Среднее значение суммы транзакции: 154,1 бразильских реала (BRL) - по одному заказу может быть несколько способов оплаты

. Все отзывы, как следует из описания данных, содержат оценку. Средняя оценка удовлетворенности по 5-ти балльной шкале - 4,1

. Средняя продолжительность доставки заказа покупателю: 12 дней

. Средняя ожидаемая продолжительность доставки заказа покупателю (сообщается ему при оформлении заказа): 23 дня

. В большинстве случаев планируемые сроки доставки соблюдаются, хотя есть и исключения

. Фактически, в январе 2018 г. продажи Olist store вышли на плато и больше не растут

. В среднем, самые высокие продажи - во вторник, понедельник уступает незначительно

. Со среды по пятницу продажи сокращаются и достигают минимума в субботу

. Самое низкое среднее значение показателя удовлетворенности - для покупок, совершенных в пятницу

. Если ограничиться периодом с января 2017 г. по август 2018 г., то средний чек составляет 172 BRL

. За исключением пика в сентябре 2017 г. значение среднего чека по месяцам различается незначительно

**Сегментирование покупателей**

Цель состоит в том, чтобы определить схожие по потребностям группы клиентов, количественно оценить их потенциальную ценность для бизнеса, разработать стратегию взаимодействия для удовлетворения этих потребностей.

Цели сегментации с точки зрения бизнес-целей Olist store:

. Увеличить количество повторных заказов

. Увеличить средний чек

. Повысить лояльность покупателей и сократить их отток

Этому должны способствовать:

. Индивидуальный подход, персонализация маркетинговых коммуникаций с учетом предыдущего опыта взаимодействия (содержание, периодичность)

. Фокусирование на наиболее выгодных для бизнеса группах покупателей

. Использование имеющихся данных для прогнозирования вероятности совершения покупок

Наиболее простой и требующий минимума информации для сегментирования клиентов способ – это RFM-анализ, предусматривающий классификацию покупателей по трем показателям: давности совершения последней покупки, частоте покупок за период анализа и сумме потраченных денег.

**RFM-анализ**

Файл olist\_store\_RFM.ipynb

Для RFM-анализа было выбрано 4 категории (хотя, скорее всего, достаточно и проще интерпретировать 3). Для определения границ сегментов был использован самый простой способ разбиения показателей – квартили.

В итоге получилось 32 сегмента.

Также предложен вариант сокращения количества сегментов до 4-ёх за счет повторения процедуры RFM-анализа.