# Отчет

## Описание данных

Датасет содержит данные о транзакциях, прошедших через сервис Amazon. Он включает такую информацию, как: InvoiceNo - номер накладной, по которой товар двигался к своему получателю; StockCode - код товара, каждый код уникален для каждого типа товара; Description - описание товара или услуги, как правило, это название товара; Quantity - количество заказанного товара; InvoiceDate - дата, когда был оплачен товар; UnitPrice - цена товара за штуку; CustomerID - уникальный ID, который присваивается тому или иному пользователю и со временем не меняется; Country - страна, из которой был заказан товар.

Стоит перейти к каждой колонке чуть подробнее. Накладные InvoiceNo содержат как числовые строки, так и те, что в начале содержат букву "С". Я заметил тенденцию, при которой большинство значений, содержащих букву "С" в InvoiceNo, имеют отрицательное значение Total, что, по моему мнению, служит не более чем меткой о возвратных заказах или каких-то иных вычетах. Например, один из вычетов имеет Quantity -1 и Description "AMAZON FEE". Если расчитать такую метрику, как стоимость заказа (Quantity умноженное на UnitPrice), то мы получаем отрицательное значение. Предположу, что это значение является вычетом Амазона - его частью прибыли со всех заказов за определенный период. Подобные вычеты отмечают буквой "С".

Это описание столбцов в наборе данных о продажах, который вероятно содержит информацию о продажах компании. В этом наборе данных присутствуют следующие столбцы:

InvoiceNo: номер счета, который может содержать как числовые, так и буквенные значения, возможно, для отслеживания разных версий одного и того же товара.

StockCode: код товара, который также может содержать как числовые, так и буквенные значения для идентификации товаров.

Description: описание товара, которое может содержать пустые ячейки, а иногда может содержать информацию о том, что произошло с товаром, например, был ли он утерян или утилизирован.

Quantity: количество товара, которое может быть положительным или отрицательным, отрицательные значения используются для вычета в отчетности.

InvoiceDate: дата заказа, которая варьируется в период с 01.12.2019 до 09.12.2020.

UnitPrice: цена товара за штуку, которая может быть только положительной или равной нулю.

CustomerID: идентификатор клиента, который может содержать пустые значения, возможно, из-за того, что это не заказы, а издержки за доставку и т.д.

Country: страна, в которой была сделана покупка, содержит только буквенные значения. Набор данных содержит информацию о продажах в 10 странах, где преобладающей страной является Великобритания с 90,7% от всех данных.

С уверенностью можно сказать, что данные качественные и более чем полные. Пустых значений чуть меньше 133 строк, если брать во внимание, то что это пустые значения из столбца Description, то такие значения можно смело отбросить и убрать из анализа.

В датасете помимо прочего встретилось аномальное значение, его я пометил как следующую строку:



Аномальная строка содержит описание: printing smudges/thrown away (пятна от печати/выброшены), вопрос о том, что означает эта строка (Возможно брак) необходимо уточнять у поставщика. Я решил избавиться от данной строки, так как она могла вносить помехи в метрики и сильно влиять на среднее.

## Подход к анализу данных

Первичный анализ проводился в самом Excel файле, применялся подход сортировки каждого столбца от А до меня и наоборот, после чего я старался найти интересные закономерности. Поняв структуру и объем данных, можно было перейти к дальнейшему анализу уже в Jupiter Notebook, где с помощью библиотеки pandasql я смог структурировать и анализировать данные используя инструменты SQLite.

Сильное ограничение накладываемое данными, с моей точки зрения это отсутствие описания датасета, какие столбцы за что отвечают, на какие данные стоит обратить внимание, и кто заказчик, позже станет ясно что вероятно это Амазон, но на ранних этапах эта информация была бы полезна.

## Описание бизнес-заказчика

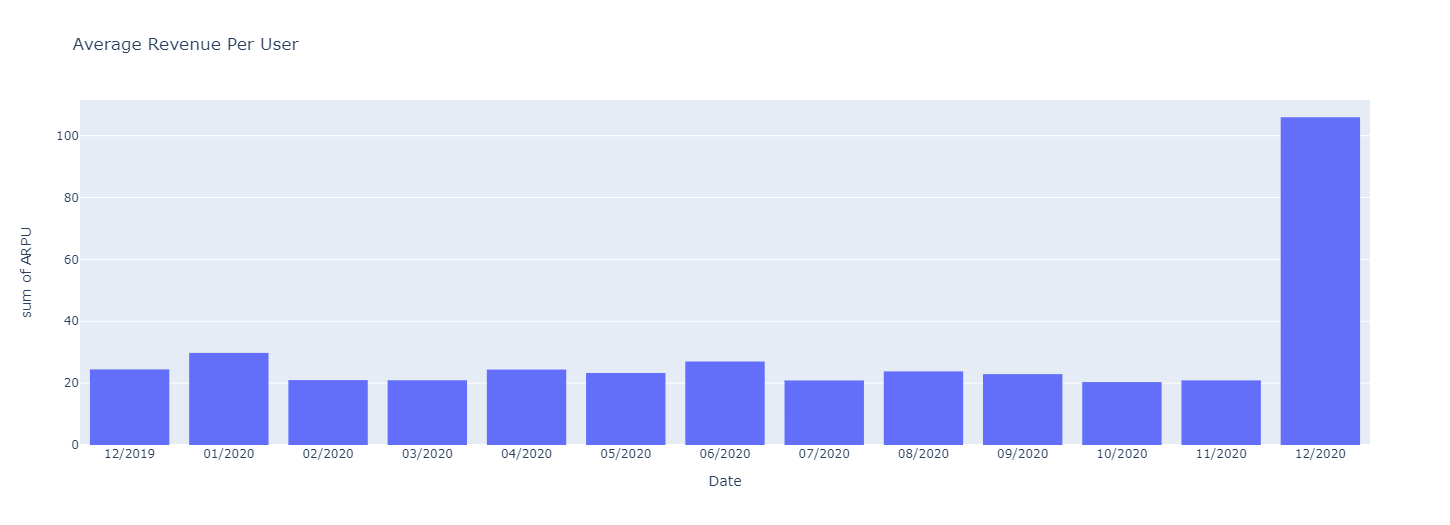
Заказчиком выступает Amazon. Предполагаю, что основной болью заказчика является удержание клиентов на сервисе (Customer Retention). Удержание клиента является огромной проблемой такого крупного сервиса, как Амазон, так как, по моему мнению, удержать клиента сервису будет проще, чем привлечь нового. Ведь об Амазоне знает по меньшей мере каждый второй только в России, а за границей о сервисе должен знать каждый.

Я смог определить процент общего оттока клиентов и процент оттока по месяцам и могу заверить, что серьезный рост оттока — это беда, от которой заказчик хотел бы избавиться. Причина возникновения роста мне не ясна. Один очевиднейший вариант, который приходит мне на ум, это задержка заказов. Хорошей практикой было бы запросить у заказчика дополнительный датасет со срезом того же времени и информацией по дате доставки заказов и фактической дате доставки заказа. Имея на руках такие данные, было бы несложно сложить 2 + 2 и подтвердить гипотезу того, что клиент, который не получил свой заказ вовремя, не сделает повторную покупку.

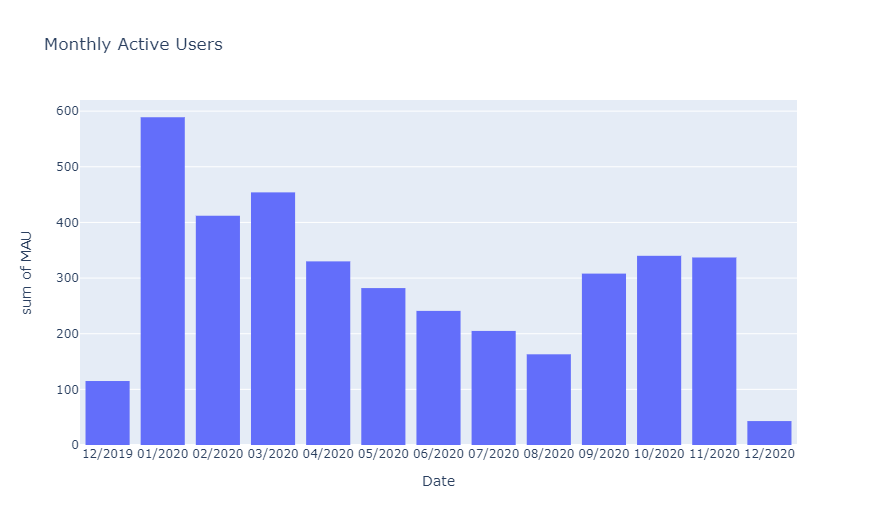
## Метрики

1. ARPU - средний доход в расчете на одного пользователя (average revenue per user). Вычисляя эту метрику, я ожидал увидеть значительный рост или падение среднего значения, но ничего подобного я не увидел кроме скачка в Декабре 2020 года, который составлял 105.957442 на одного пользователя, в то время как максимальный рост за другие месяца составляет 29.773664. Такой скачок обусловлен низким количеством клиентов, датасет резко обрывается и 43 пользователя заказавших на крупные суммы дали о себе знать.

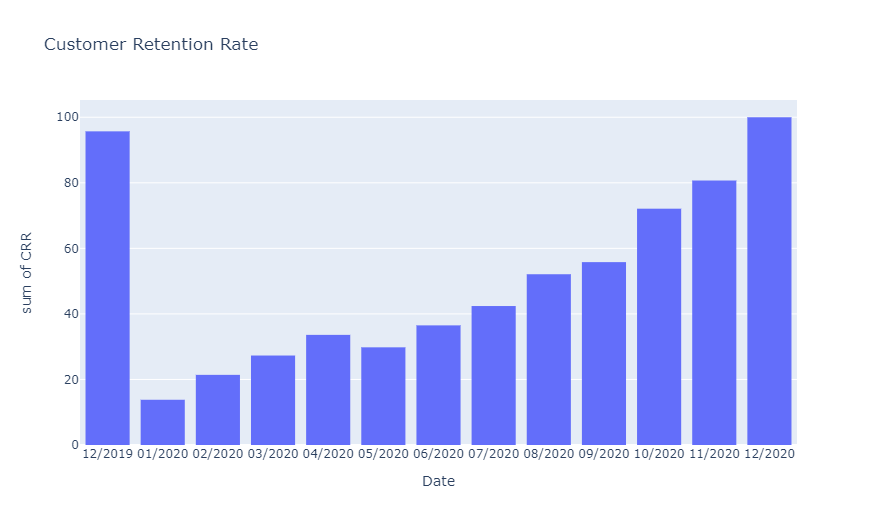
Гистограмма среднего дохода в расчете на одного пользователя.



1. MAU - Число активных пользователей в месяц (Monthly Active Users). Метрика необходимая для вычисления другой, основной метрики. Как можно видеть из гистограммы ниже пиковая активность приходилась на Январь 2020 года и составляла 589 активных пользователей.



1. Следующая метрика имеет общие данные по датасету, она не несет никакого практического применения, а выступает как отчетность. Общее количество накладных составляет 13673, уникальных пользователей за весь отчет 3819. Общая сумма всех заказов 840206.211. Среднее значение стоимости заказа 20.672839 и средняя стоимость заказа в накладной 61.450026.
2. CRR - Рейтинг оттока клиентов (Customer Retention Rate). Метрика показывает какой процент клиентов, от общего значения клиентов, за целый месяц решил покинуть сервис и соответственно не совершать повторную покупку. Гистограмма ниже указывает на большой уровень оттока клиентов с сервиса. На Декабрь 2020 отток составляет 100% по причине обрыва датасета с минимальным количеством значений за Декабрь.



1. RCR - процент покупателей, сделавших повторную покупку (Return Customer Rate). Имея данную метрику, я смог вычислить следующие значения: Всего имеется 3819 уникальных клиентов, из них 2229 сделали повторную покупку, 1590 клиентов сделали заказ один раз и больше не возвращались к сервису за период отчета. Следовательно, Return Customer Rate = 58.37%.

## Рекомендации

Основную рекомендацию я представляю следующим образом: Сделать распределение всех моих метрик дополнительно по странам и отслеживать страну, в которой возможно было бы увеличить процент активных пользователей, благодаря которым можно было бы развернуть реферальную маркетинговую компанию. Благодаря такой компании мы могли бы увеличить привлечение клиентов на сервис, если, конечно, это позволяет конкурентоспособность.

## Вывод

Как вывод, сервис терпит огромный отток клиентов. Чтобы спастись от этой ситуации, следует капнуть глубже, получить дополнительные данные, которые позволят нам ответить на вопрос: Почему? Помочь нам узнать все нюансы и углубиться в ситуацию может только заказчик. Найти спасение через сторонние сервисы, предоставляющие данные, мы не можем, так как вся информация по заказам, их доставке, адресатах и прочих конфиденциальных данных принадлежит заказчику.