Онлайн университет Нетология

**Группа DA-46**

**ДИПЛОМНАЯ РАБОТА**

**"Анализ продаж британского e-commerce (поиск инсайтов, составление рекомендаций стейкхолдерам, построение предиктивной модели объёмов продаж)"**

Профессия: Аналитик данных

Шакула Вадим Анатольевич

Москва 2022

Содержание

1. Постановка задачи ……………………………………………………………………………………………………….3
2. О наборе данных ………………………………………………………………………………………………………….3
3. Система метрик …………………………………………………………………………………………………………….3
4. Рекомендации по модели данных ………………………………………………………………………………4
5. Описание последовательности действий в colab …………………………………………….…………4
   1. Загрузка библиотек и открытие файла ……………………………………………………………………4
   2. Разведочный анализ данных …………………………………………………………………………………..4
   3. ДФ для предварительной аналитики (поиск топ в sku, стране, пользователях) …..6
   4. ДФ для аналитики и визуализации (основа) …………………………………………………………..8
   5. Прогнозирование …………………………………………………………………………………………………….13
      1. Линейная регрессия …………………………………………………………………………………………13
      2. RMSE среднеквадратичная ошибка …………………………………………………………………14
      3. Вывод RMSE ………………………………………………………………………………………………………14
6. Заключение …………………………………………………………………………………………………………………..15

**Постановка задачи**

**Бизнес – задача** – провести аналитику данных по продажам, подготовить данные к предварительной аналитике (убрать данные, которые могут привести к неверным выводам) и подготовить визуализацию ключевых показателей. Спрогнозировать продажи будущих периодов.

Основная цель компании по продажам – увеличение дохода компании. Стратегия определить основные метрики для определения эффективности продаж, так и маркетинговых активностей.

**Стейкхолдерами** аналитики данных выходят отдел продаж (для своевременного реагирования на изменения продаж), маркетологи (для оценки влияние рекламы на покупательскую способность, трейд-маркетологи (для оценки эффективности проводимого промо и своевременной корректировки для увеличения ROI), руководительский состав (для оценки экономических показателей компании).

**О наборе данных**

[[1]](#footnote-1)"Это транснациональный набор данных, который содержит все транзакции, произошедшие между 01/12/2010 и 09/12/2011 для зарегистрированной в Великобритании онлайн-розничной торговли, не связанной с магазином. Компания в основном продает уникальные подарки на все случаи жизни. Многие клиенты компании являются оптовиками."

Наименование полей и их расшифровка:

1. InvoiceNo: номер транзакции, состоит из 6 цифр.
2. StockCode: код товара (позиции).
3. Description: описание товара
4. Quantity: количество товара
5. InvoiceDate: дата транзакции
6. UnitPrice: цена товара за единицу
7. CustomerID: номер клиента. Уникальное 5-значное целое число, присваиваемое каждому покупателю.
8. Country: название страны, в которой проживает каждый покупатель.

Данные содержат 54214 строк, есть пропуски – description содержит 54080 ненулевых объектов, customerID – 40643 ненулевых объекта. Для качественной аналитики не хватает некоторого количества описаний транзакций, что продали и кому.

Quantity может быть отрицательной – это возвраты товаров клиентами, если CustomerID заполнено. Если CustomerID пуст совместно с Description, но при наличии инвойса это говорит об ошибке в системе интернет-магазина.

Для проведения аналитики желательно заполнить данные или исключить их при дальней аналитике и построения отчетности.

**[[2]](#footnote-2)Система метрик могут выходить показатели:**

Брошенные корзины

Процент возврата пользователей

Показатель отказов

Стоимость действий

Кол-во штук в одном заказе

Стоимость 1 штуки в одном заказе

Средний чек

Макро и микро-конверсии

**Рекомендации по модели данных бизнес-заказчика**

Для анализа и получения доп. метрик потребуется так же информация по клиентам (данные возраста, пола, год рождения, даты регистрации, с какого сайта перешел, но также для расчета эконом показателей потребуется цена закупки сырья, логистика, себестоимость. Возможно, расчет EBITDA , такой показатель нужен инвесторам) – *запрашиваемые данные носят рекомендательный характер для того, чтобы покрыть полный спектр потребности внутренних и внешних стейкхолдеров*.

**[[3]](#footnote-3)Описание последовательности действий (краткий чек-лист описан в конце проекта в colab):**

**Загрузка библиотек и открытие файла**

1. Загрузка библиотеки NumPy и pandas.



1. Загрузка файла с применением кодирования ISO-8859-1.

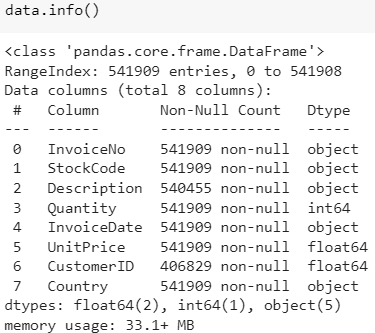


**Разведочный анализ данных**

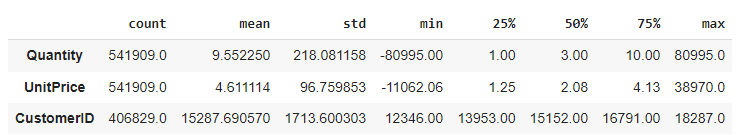
1. Первично оцениваем колонки, количество строк и тип данных.



1. Анализируем данные на наличие пустот в них.



1. Смотрим статистику по числовым данным.



1. Удаляем строки с отрицательными значениями в прайсе.





1. Смотрим минусы в количествах (Quantity), принимаем решение их оставить.



1. Смотрим на наличие пустот в кастомерах (CustomerID), заполняем её “0”.





1. Проверяем товары (Description), все что без описания удаляем из данных. Такие данные могут повлиять на аналитику в дальнейшем при прогнозировании.



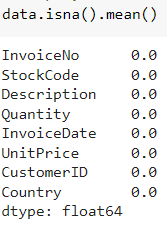




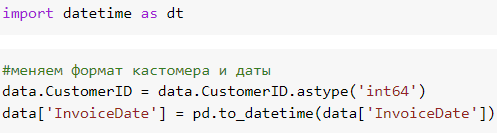
1. Проверяем цену по прайсу (UnitPrice), все товары с “0” ценой удаляем.



1. Проверяем информацию по датафрейму, все ли пустоты заполнены. Вся пустота заполнена.



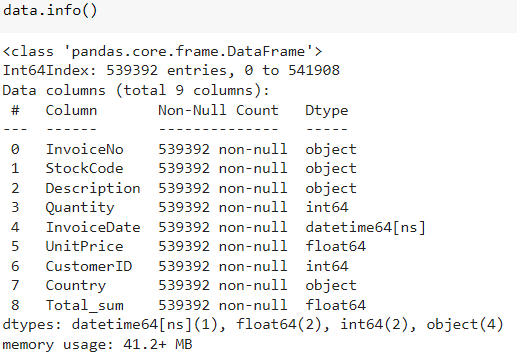
1. Меняем тип данных для кастомеров и даты.



1. Добавляем расчетное поле Тотал\_сумма, которое рассчитывается как кол-во\*прайс. Выводим первую ключевую метрику, которую будем также использовать для дальнейших вычислений и построения графиков.



1. Окончательно проверяем ДФ на наличие всех данных, типов данных и добавленного поля.



**Выводы развед анализа данных:**

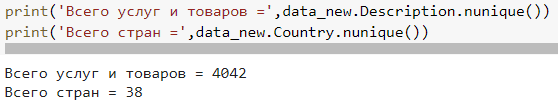
предварительная обработка завершена- пустота заполнена, сомнительные инвойсы удалены, формат данных кастомер и дата изменен, данные, которые могут повлиять на планирование удалены. Можем продолжать анализировать.

**Создаем ДФ для предварительной аналитики (поиск топ в sku, стране, пользователях)**

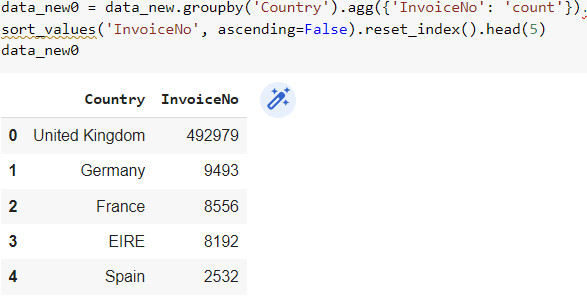
1. Создаем новый ДФ с условием, где все цены в прайсе больше 0.



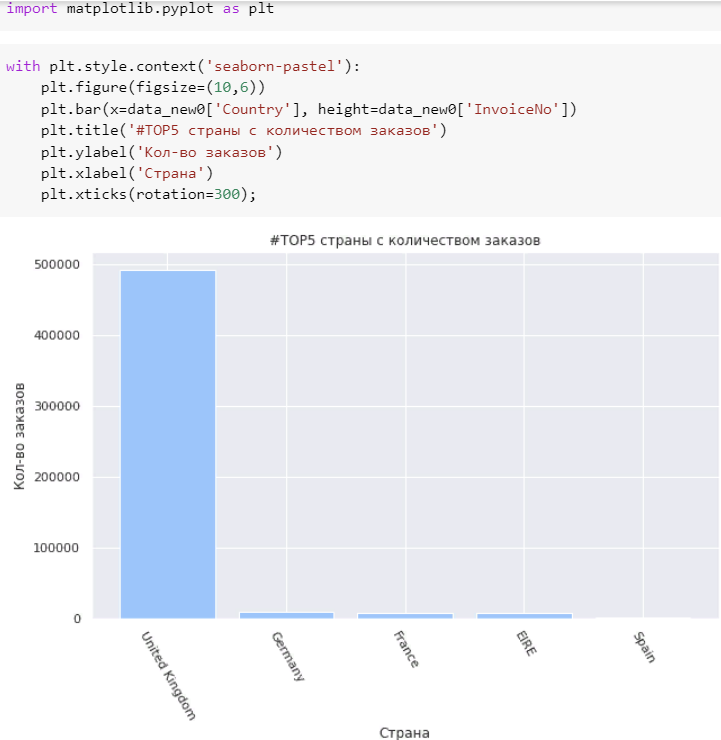
1. Оцениваем количество товаров и стран.

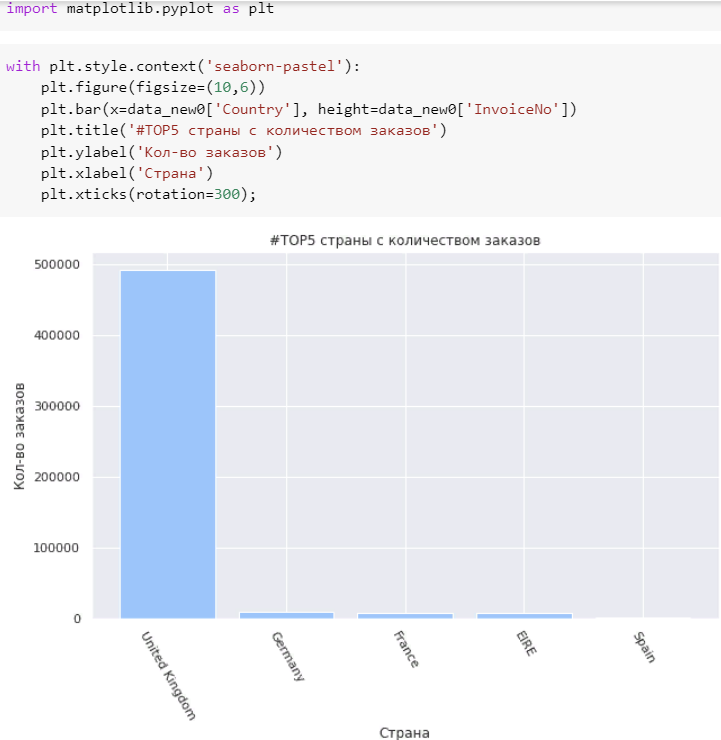


1. Считаем ТОП по странам с максимальным количеством заказов. Оценим, какая из стран делает больше всего заказов.



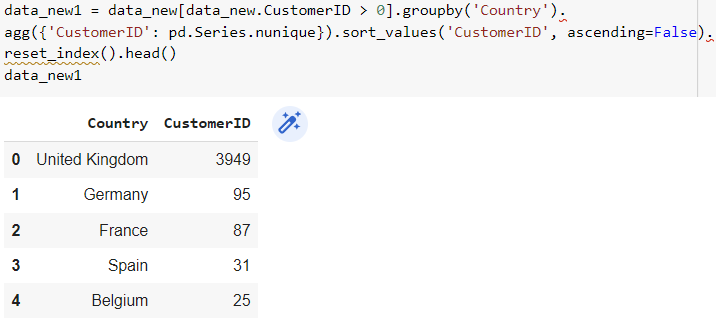
1. Строим гистограмму для визуализации данных.



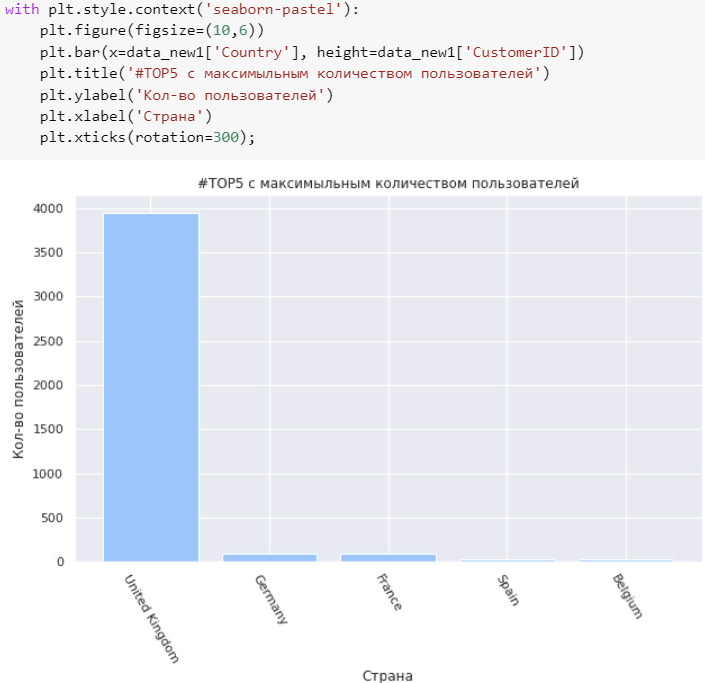


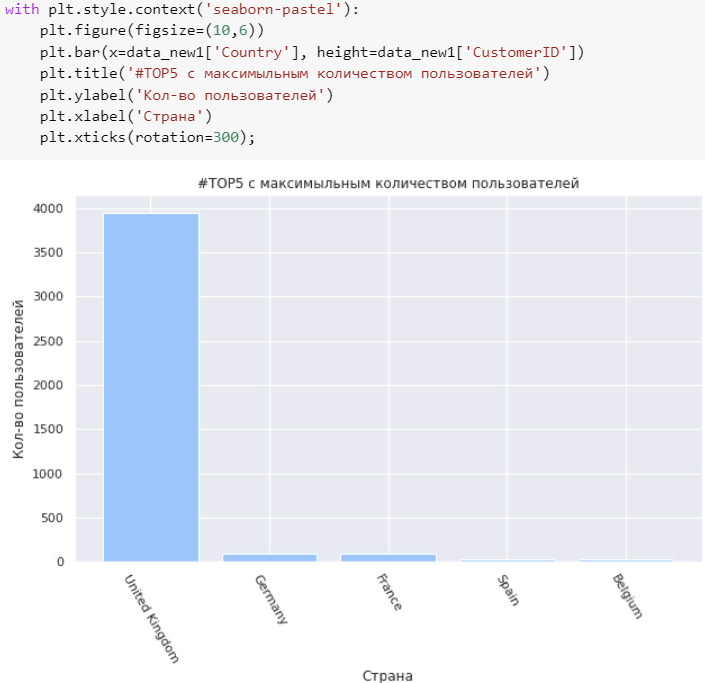
**Вывод:** Топ по странам занимает United Kingdom c числом заказов 492979, следующая страна по численности Germany – 9493 заказа за период.

1. Считаем ТОП странам по количеству пользователей. Определение лидера по числу пользователей.



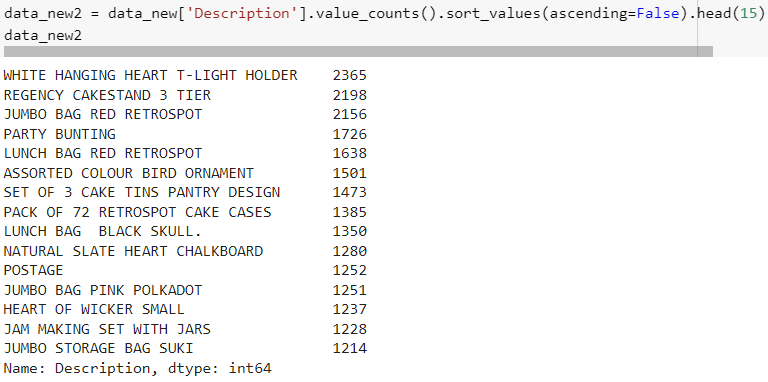
1. Строим гистограмму для визуализации данных.





**Вывод:** Топ по странам занимает United Kingdom c числом пользователем 3949, следующая страна по численности Germany – 95 пользователей.

1. Считаем ТОП позиций по заказам.



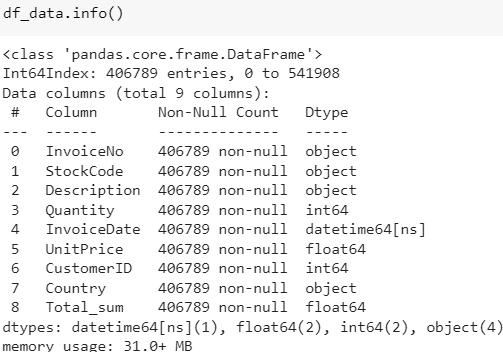
**Вывод** Самый заказываемый продукт WHITE HANGING HEART T-LIGHT HOLDER

**Создаем ДФ для аналитики и визуализации**

1. Создаем ДФ в которых убираем всех кастомерах (CustomerID) значение которых меньше “0”.

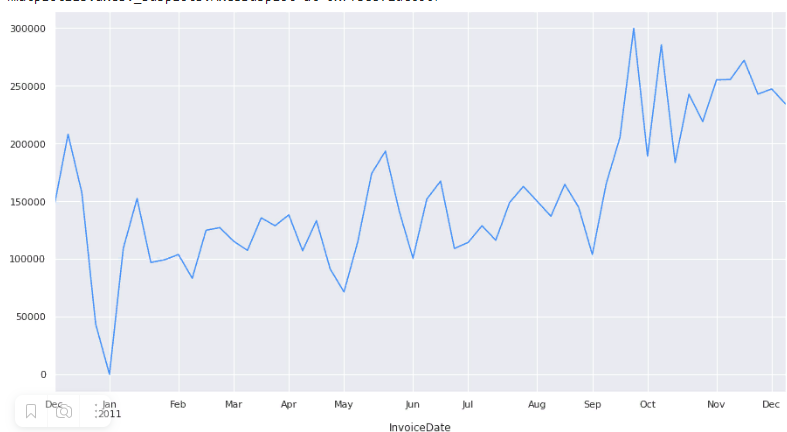


1. Смотрим данные через info().



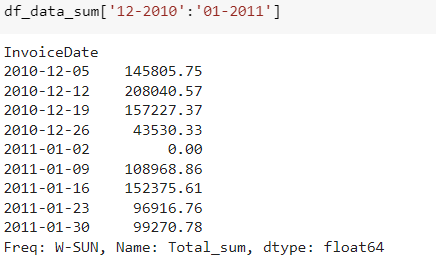
1. Смотрим выручку. Строим визуализацию.

Для этого создаем дополнительный дф, добавляем период год-мес и суммируем продажи по периодам.



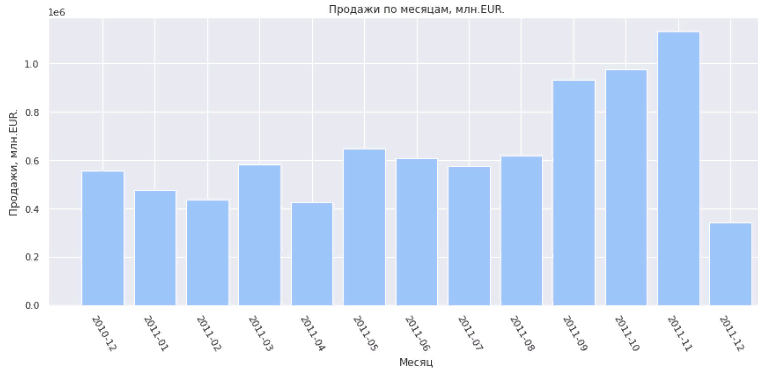
Отклонение, которое требует проверки. Замечаем странное, в январе падение. нужно выяснить почему.

1. Проверяем период продаж с 12-2010 по 01-2011. Более глубокая аналитика даст нам ответ почему в начале января продажи равны “0”.



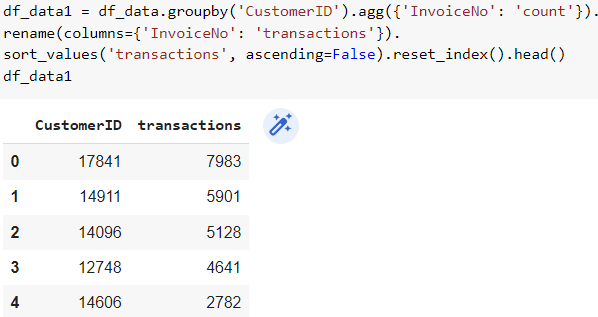
нет продаж за 1.02 , в данный период был выходной день.

1. Строим визуализацию по месяцам с продажами, чтобы оценить на каком этапе сейчас компания (рост, стагнация или падение)



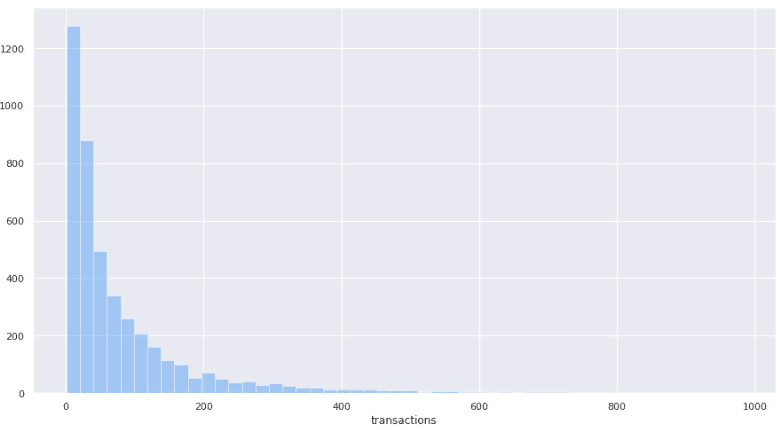
**Вывод:** на графике выше видим, как распределяются продажи относительно каждого месяца. Наибольший пик наблюдается в период с октября по ноябрь. Возможно, это связано с тем, что люди массово покупают подарки к Рождеству и Новому году. Вместе с тем отмечается резкий спад продаж в конце декабря 2010 г. - январе 2011 г. Вероятнее всего в период Новогодних праздников магазин не работает.

1. Смотрим **ТОП по пользователям и их транзакциям.**



Данные позволят выявить кто из пользователей совершил больше всего покупок за указанный период и принять решение о возможном поощрении (программы лояльности, накопительная/бонусная карты, скидка постоянного клиента).

1. Смотрим **какое количество заказов делают пользователи.**



**Вывод** исходя из визуализации выше, что большинство клиентов (более 1000 человек) совершили до 20 транзакций

**Смотрим количество уникальных пользователей за каждый месяц**

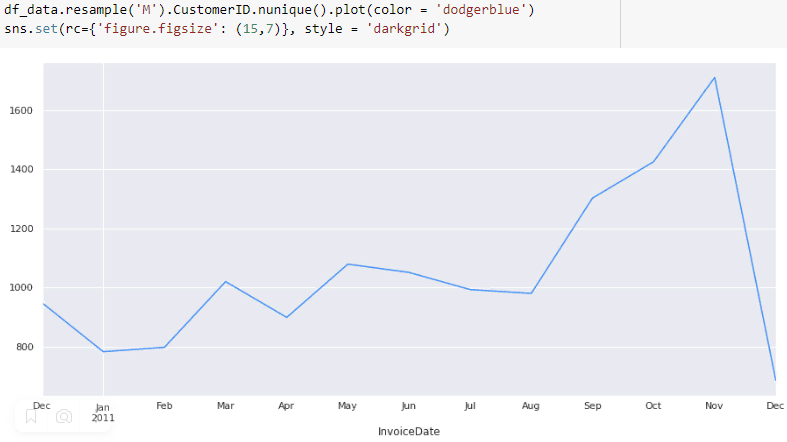
1. Добавляем дату в ДФ.



1. Проверяем ДФ, добавлена ли дата в него.



1. Делаем визуализацию в размере месяца и количества уникальных пользователей и делаем выводы для ответственных.



**Вывод** Максимальный месяц с уник пользователям ноябрь, свыше 1600 пользователей сделали заказ. Скорее всего это связано с предновогодним спросом, когда компания увеличила активность в маркетинге.

**Смотрим количество транзакций за каждый месяц. Для этого строим визуализацию.**

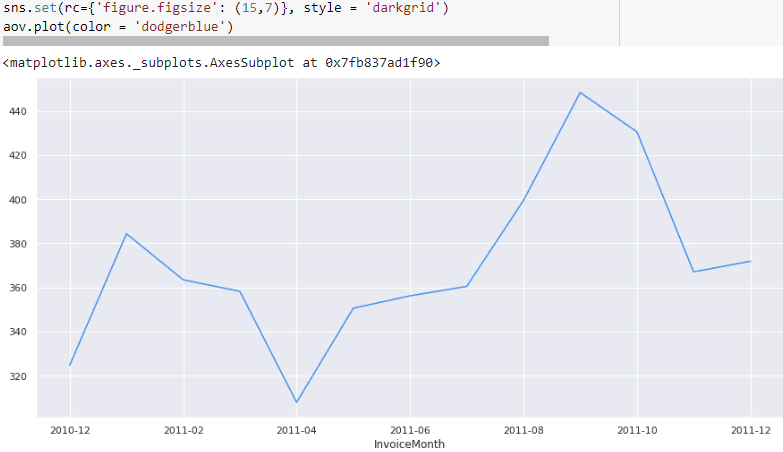


**Вывод** Максимальное количество транзакций ноябрь. Это подтверждает вывод выше — это связано с предновогодним спросом.

**Смотрим информацию по среднему чеку (AOV).**



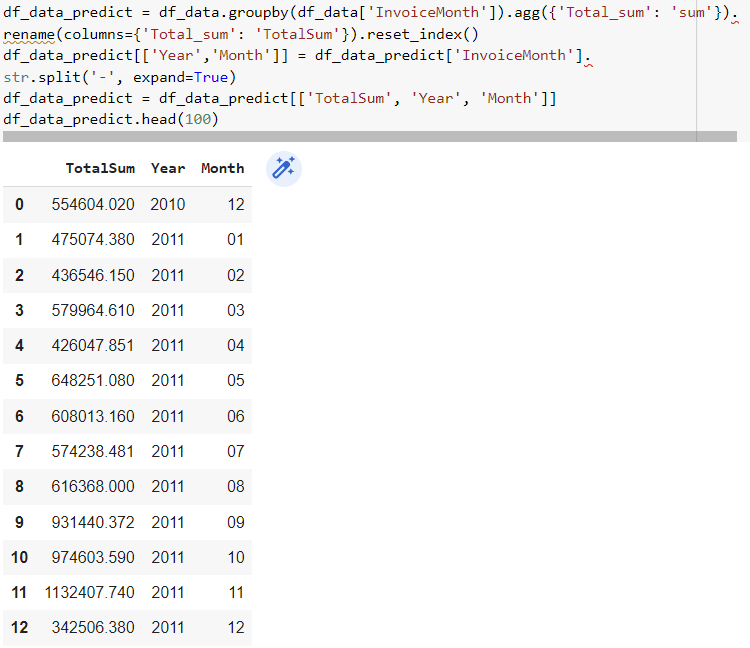


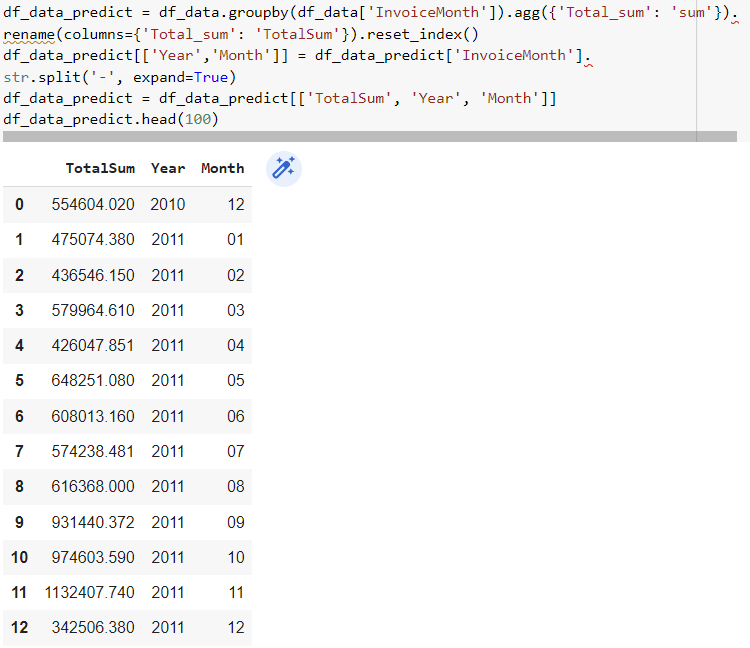
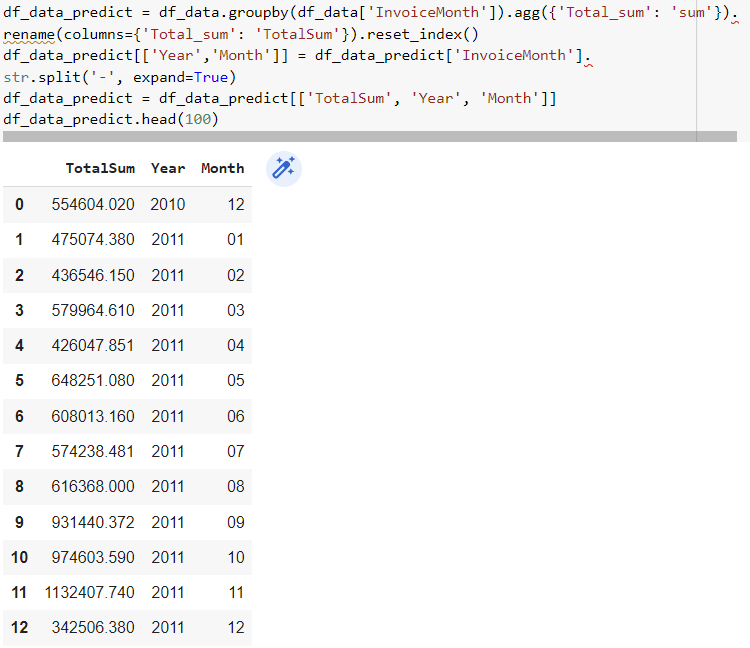


**Вывод.** Из визуализации видно, что максимальный средний чек (AOV) был в сентябре, но в свою очередь количество заказов и пользователей упало. (это видно из графиков выше).

**Прогнозирование**

1. Подготовим данные для дальнейшей работы. Создаем дф в котором у нас будут месяц, год и тотал продажи за период.

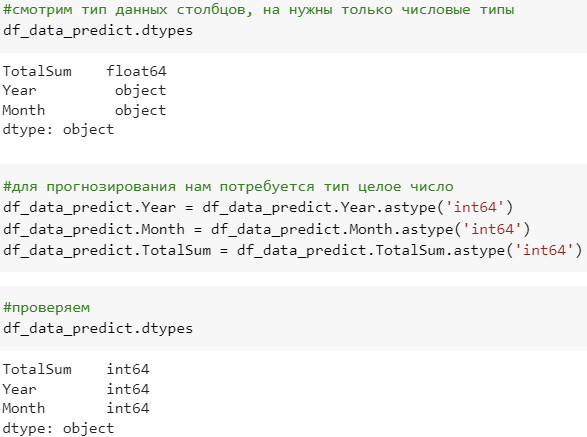


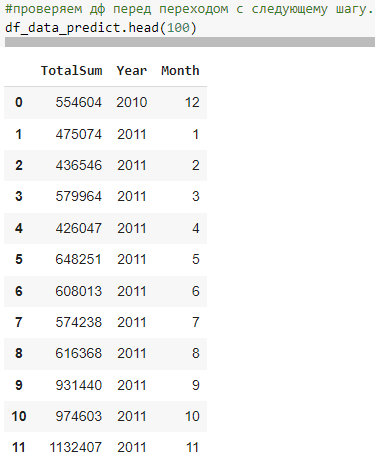
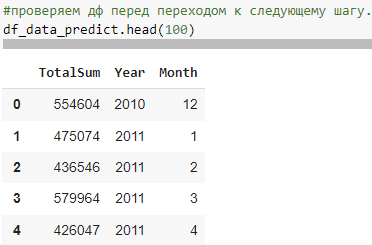


1. Удаляем 12-й месяц 2011 года из ДФ, так как он не полный. Он влияет на предсказание продаж.



1. Переводим типы данных в целое число



****

**Линейная регрессия**

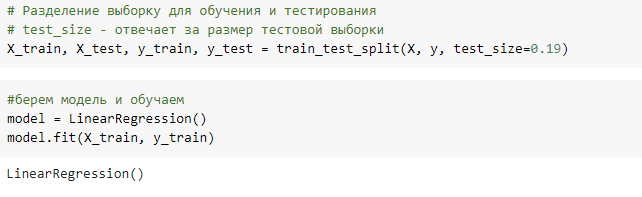
1. Импортируем библиотеки.



1. Признаки

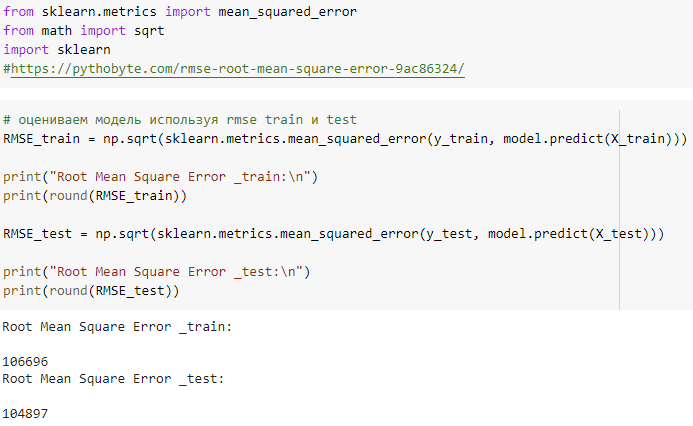


1. Разделяем выборку, берем модель и обучаем.

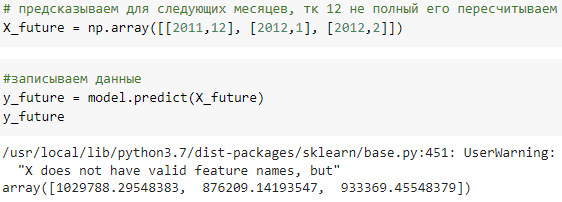


**[[4]](#footnote-4)RMSE среднеквадратичная ошибка**

1. Импортируем библиотеку и оцениваем модель.



1. **Прогнозируем** продажи на период 12.2011 – 03.2012.



**ВЫВОД по RMSE и прогнозу будущих периодов:**

**RMSE для train и test наборов очень похожими (train > test), это говорит о том, что мы построили хорошую модель. Модель имеет большую прогностическую ценность.**

**RMSE\_train: 106 696**

**RMSE\_test: 104 897**

**В этом случае мы можем доверять прогнозу, произведенным после, но нужно учитывать обучения и оценки модели погрешность (которая составляет ~105 000).**

**12.2011 - 1 029 788**

**01.2012 - 876 209**

**02.2012 - 933 369**

**Заключение**

Согласно метрике AOV видим, что месяцем с наиболее высоким показателем среднего чека является сентябрь. Показатель AOV можно увеличивать за счет продажи сопутствующих товаров. Как пример, при продаже верхней одежды можно предлагать клиентам купить водонепроницаемый чехол для одежды из инновационной сверх дышащей ткани. Стимулировать рост среднего чека для компании может быть выгоднее, чем привлекать новых клиентов. Готовому к покупке клиенту не нужно еще раз знакомиться с компанией, заново читать отзывы других покупателей или гуглить преимущества товара этого бренда.

Наибольшее количество покупок совершено в октябре - ноябре, то есть перед праздниками. Максимальное число вновь пришедших клиентов приходится как раз на этот же период. Разумеется, фактор сезонности нельзя исключать, но для того, чтобы продажи распределялись более-менее равномерно в течение всего года, маркетологам следует уделить внимание поощрению наиболее лояльных клиентов, возможно также сделать акцент на сезонные акции.

Сгруппировав наши данные по странам и выявив количество уникальных пользователей, мы обнаружили, что более 90% покупателей приходится на Великобританию. Таким образом, можем заключить, что если руководство компании видит своей целью увеличение объемов продаж, то выход на новые рынки, либо укрепление на уже существующих, однозначно необходим. Возможно маркетологам следует дополнительно проанализировать работу каналов привлечения трафика: какие из них дают наибольшую конверсию.

1. [E-Commerce Data (About Dataset)](https://www.kaggle.com/datasets/carrie1/ecommerce-data?datasetId=1985&sortBy=dateRun&tab=collaboration) [↑](#footnote-ref-1)
2. [Глоссарий Нетология](https://netology.ru/glossariy) [↑](#footnote-ref-2)
3. [Ноутбук дипломной работы](https://colab.research.google.com/drive/1KFZ0WPP1Cc7lxT9dCWIDoSeY0Jpkqnn9?usp=sharing) [↑](#footnote-ref-3)
4. [Rmse – root средняя квадратная ошибка в Python](https://pythobyte.com/rmse-root-mean-square-error-9ac86324) [↑](#footnote-ref-4)