**Пояснительная записка к итоговой работе**

**по курсу «Аналитик данных»**

**На тему: «**Анализ продаж e-commerce (исследование данных, поиск инсайтов, составление рекомендаций стейкхолдерам, построение модели)»

**Лазарева Елена Владимировна**

**Группа: DA-19**

**2021 г.**

**Оглавление**

[**Введение** 3](#_Toc78719752)

[**Постановка** **задачи** 3](#_Toc78719753)

[**Анализ датасета** 4](#_Toc78719754)

[**1.** **Посмотрим на датасет** 4](#_Toc78719755)

[**2.** **Удаление пропусков и неадекватных значений** 4](#_Toc78719756)

[**Проведение расчетов** 5](#_Toc78719757)

[**ABC-XYZ-анализ** 9](#_Toc78719758)

[**RFM-анализ** 12](#_Toc78719759)

[**Вывод** 15](#_Toc78719760)

# **Введение**

В итоговой работе проанализирован датасет [E-Commerce Data](https://www.kaggle.com/carrie1/ecommerce-data/home) с данными по онлайн-продажам с сайта [kaggle.com](https://www.kaggle.com/).

Это транснациональный набор данных, который содержит все транзакции, произошедшие с 01/12/2010 по 09/12/2011 для британской компании онлайн-розничной торговли. Компания в основном продает уникальные подарки на все случаи жизни. Многие клиенты компании являются оптовиками.

**Описание переменных в датасете:**

* **InvoiceNo:** номер счета, состоящий из 6 цифр. Если этот код начинается с буквы «c», это означает отмену;
* **StockCode:**код продукта, состоящий из 5 цифр;
* **Description:** название продукта;
* **Quantity:** количество каждого продукта за транзакцию;
* **InvoiceDate:** представляет день и время создания каждой транзакции;
* **UnitPrice:** цена продукта за единицу;
* **CustomerID:** номер клиента, состоящий из 5 цифр. У каждого клиента есть уникальный идентификатор клиента;
* **Country:** название страны, в которой проживает каждый покупатель.

# **Постановка** **задачи**

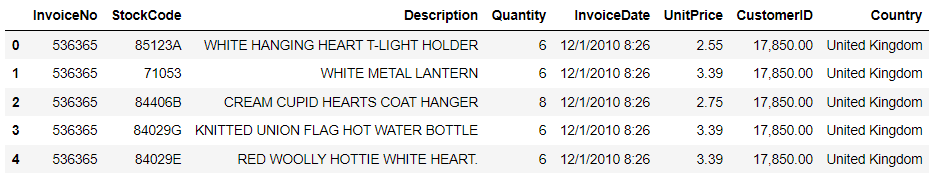
Основная цель исследования – анализ продаж. Данные по продажам позволяют:

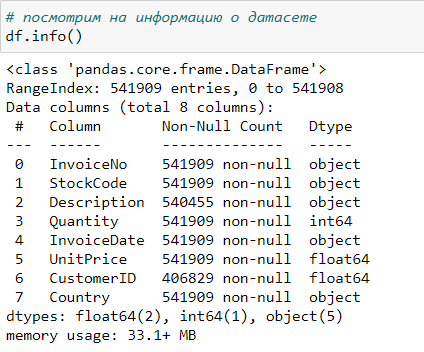
* рассчитать выручку и провести ее анализ,
* провести анализ клиентов;
* исследовать фактора сезонности в разрезе месяцев по визуализации объёмов продаж.
* провести RFM-анализ для разработки релевантных коммуникаций и контента. Лояльным клиентам — спецпредложения. Тем, кто давно не покупал — бонус или скидку + таргетинг, чтобы напомнить о себе.
* провести ABC-анализ, что позволяет выделить приоритетную группу по количественным показателям и сосредоточить усилия на работе с ними.
* спрогнозировать выручку на ближайшие 3 месяца.

Регулярный анализ продаж позволяет Департаменту продаж своевременно оценить нужно ли обновить ассортимент или активизировать работу с потенциальными клиентами, что позволит определить дальнейшую стратегию развития продаж.

# **Анализ датасета**

1. **Посмотрим на датасет**



****

Датасет содержит 8 колонок и 541 909 строк. Типы данных, содержащиеся в датасете: float64, int64 и object.

Столбец с датой InvoiceDate имеет тип данных object, который не предназначен для работы с датами. Преобразуем тип данных в данном столбце в datetime. Столбец CustomerID преобразуем к целочисленному, т.к. это идентификатор.

1. **Удаление пропусков и неадекватных значений**

В столбце Description (Название продукта) содержится 1 454 пропуска, а в столбце CustomerID (Номер клиента) 135 080 пропусков. Строки с пропусками составляют 25% от всего датафрейма.

Всего 135 080 строк, где есть пропуски и в описании товара, и в идентификаторе пользователя. Также в данных строках наблюдаются отрицательные и нулевые значения цены. Объяснение этих транзакций с пропусками в данных отсутствует и лучше такие строки исключить.

Также наблюдаем отрицательные значения количества товаров и нулевые значения цены товара. Можно предположить, что строки с отрицательными значениями количества товаров - возвраты товаров. Возвраты товаров нам не интересны. У нас также отсутствует объяснение происхождения записей с нулевой ценой. Удалим данные строки из датафрейма, т.к. они не несут нам полезной информации.

# **Проведение расчетов**

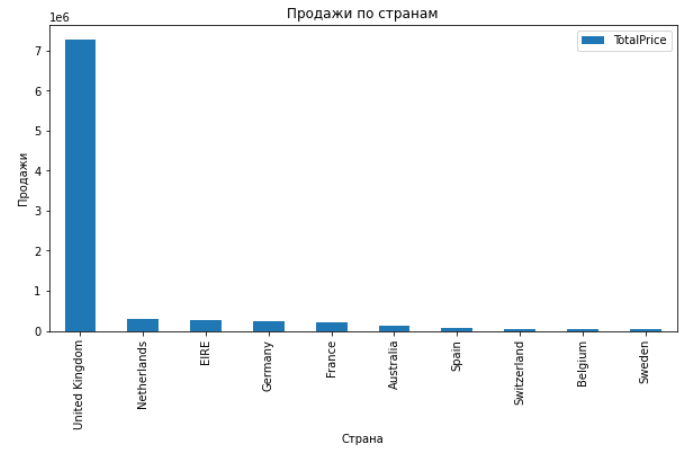
В датасете есть данные только по розничной цене товара и количеству товара. В новый столбец TotalPrice добавим итоговую выручку по каждому товару (кол-во \* цену за единицу).

Посмотрим на корреляционную зависимость между переменными. Наблюдаем, что наиболее скоррелированы между собой переменные TotalPrice и Quantity (сильная прямая зависимость). Остальные переменные имеют слабую зависимость между собой.



Выручка компании за период с 01/12/2010 по 09/12/2011 составила 8 887 208.894.

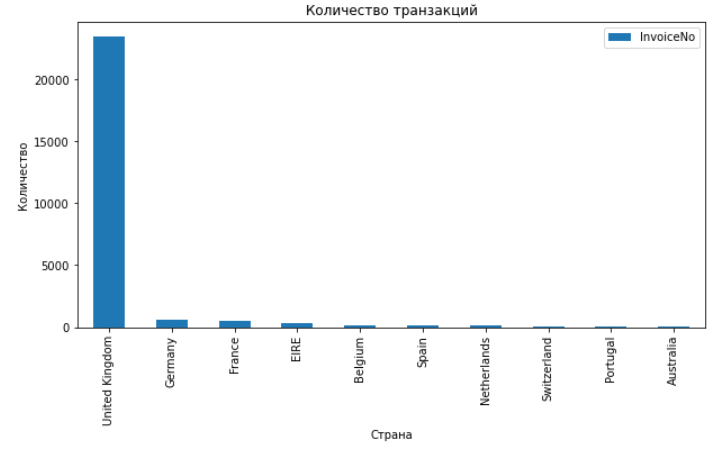
Видим, что наиболее высокая выручка в Великобритании. Она существенно выше, чем выручка в остальных странах.



Можно предположить, что продажи совершались не во всем анализируемом периоде, т.к. Компания вышла на международный рынок недавно, либо Компания, наоборот, прекратила продажи в других странах.

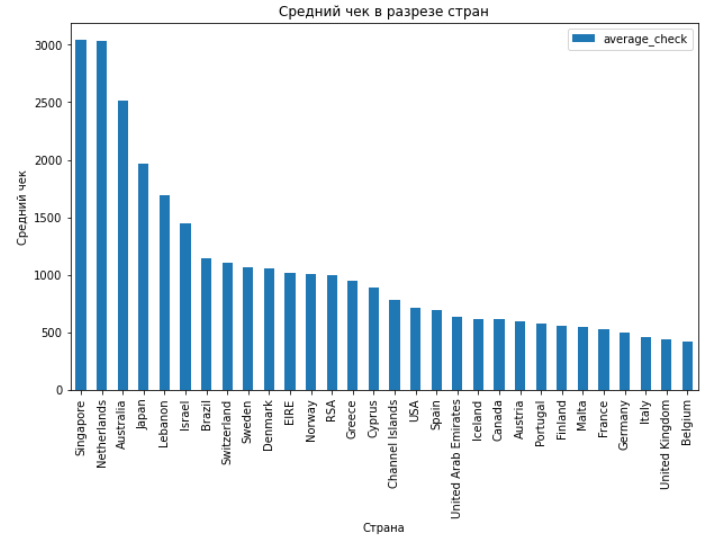
Проверим данную гипотезу. Найдем дату самой ранней транзакции, совершенной за пределами Великобритании – 01.12.2010 и дату самой поздней транзакции – 09.12.2011. Гипотеза не верна. Скорее всего, Компания не особо известна или популярна в других странах. Компании можно развивать продажи и повысить узнаваемость бренда в международном направлении.

Великобритания имеет также наиболее количество транзакций.

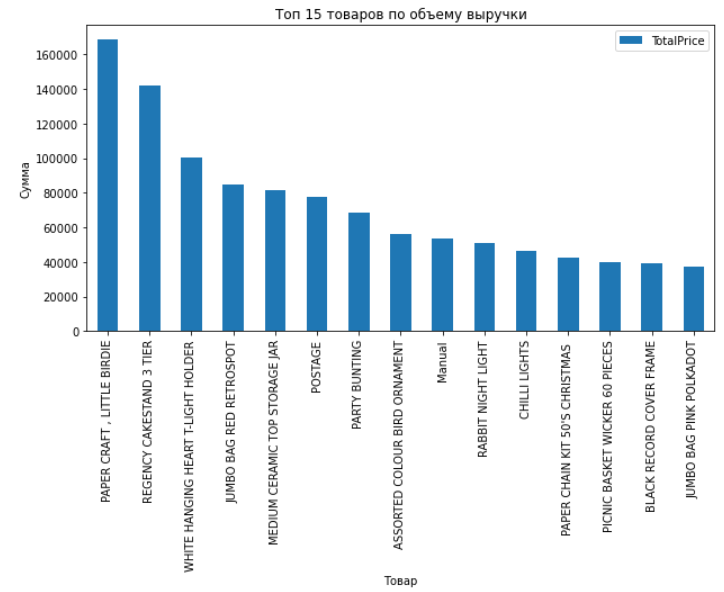


Посмотрим на средний чек в разрезе стран. Самый высокий средний чек в Сингапуре, Нидерландах и Австралии. Это может быть объяснено тем, что в данных странах было совершено немного транзакций, но на большую сумму

В Великобритании, где сосредоточены наиболее высокие продажи, наоборот же, средний чек составляет всего 438.80



Посмотрим на товарные позиции, которые приносят наибольшую выручку. Т.к. данные товары наиболее продаваемые, то необходимо иметь их запас на складе.



Оценим сезонность продаж. Продажи в Декабре 2010 и Январе 2011 года примерно на одном уровне. С Февраля по Май наблюдаем нестабильность продаж. С Мая по Август отсутствуют резкие скачки, продажи примерно на одном уровне. С Августа наблюдается рост продаж. В декабре 2011 продажи слабые, но это обусловлено тем, что у нас есть данные только до 9 декабря 2011. Рост продаж ближе к концу года объясняется тем, что многие начинают покупать подарки к Новогодним праздникам.



# **ABC-XYZ-анализ**

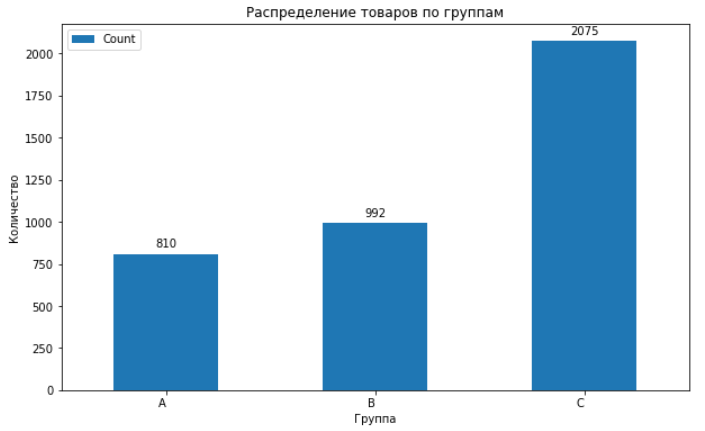
ABC-XYZ-анализ позволяет понять, какие товары приносит основной доход, а от чего лучше отказаться.

**Цель данного анализа:** выявить в ассортименте товары, которые продаются плохо и определить дальнейшую стратегию по работе с данным товаром.

Разобьем все товары на три группы:

* А — дают 80% выручки;
* В — формируют 15% выручки;
* С — дополняют оставшиеся 5%.

Видим, что у нас 810 товарных позиций (далее – ТП), которые дают на 80% выручки, 992 ТП, формирующие 15% выручки, и подавляющее большинство ТП (2075 ТП), которые дают всего 5% от выручки.



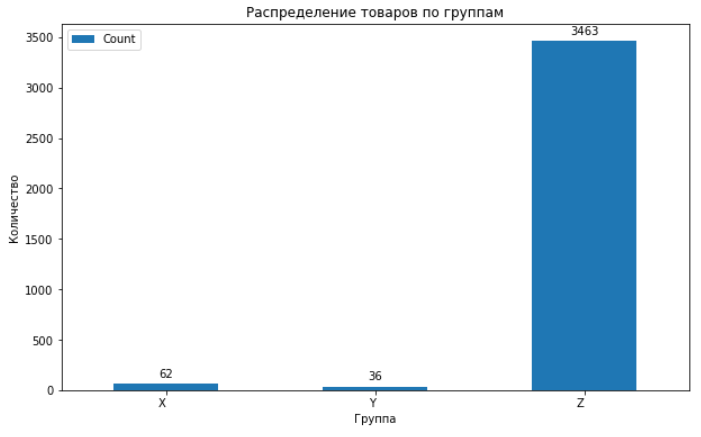
XYZ-анализ распределяет товары в зависимости от стабильности спроса.

Для распределения товаров по группам X, Y и Z необходимо:

1. **определить среднемесячную выручку**
2. **рассчитать стандартное отклонение**
3. **рассчитать коэффициент вариации (КВ).** Чем меньше коэффициент вариации, тем стабильнее спрос на товар.

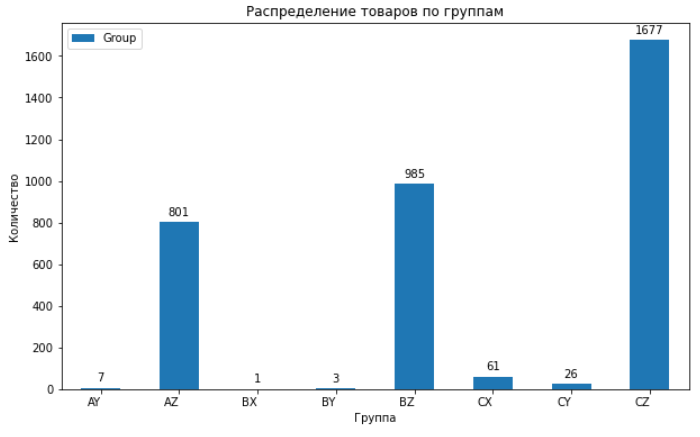
Зная коэффициент вариации по каждой ТП разбиваем на 3 группы:

* Группа Х — стабильный спрос, объем продаж от месяца к месяцу почти не изменяются: КВ < 0,1.
* Группа Y — условно стабильный спрос, т. е. объемы продаж меняются, но без резких скачков: 0,25 > КВ > 0,1.
* Группа Z — спрос нестабилен, покупатели приобретают товар от случая к случаю: КВ > 0,25.

ТП со стабильным спросом и условно стабильным спросом очень мало (62 и 36 ТП соответственно). Зато 2463 ТП имеют нестабильный спрос. 

Объединим результаты ABC и XYZ анализов и выделим 9 групп товаров:

* **Товары из групп AX и BX** (**1 ТП**) обеспечивают основной доход и при этом стабильно продаются. Они всегда должны быть на складе. Необходимый запас легко спрогнозировать, т.к. объемы продаж почти не меняются от месяца к месяцу.
* **Группы АY и BY (10 ТП)** - тоже высокий доход, но стабильность продаж уже ниже. По этим товарам на складе необходимо создать дополнительные резервы, на случай роста спроса в отдельные периоды. Резерв нужен обязательно, это важные для компании группы, которые приносят значительную часть дохода.
* **Группы AZ и BZ** (**1 786 ТП**) тоже приносят существенный доход. Но спрос на них почти невозможно спрогнозировать. Если создавать запасы на складе, исходя из максимально возможной месячной выручки, то значительная часть товара может остаться непроданной. Поэтому лучше использовать другие способы, которые обеспечат нужное количество товара.
* **Группа CX (61 ТП)** продается стабильно, но в небольших объемах. По этой позиции нужно создать постоянный запас, исходя из среднего объема продаж.
* **Группа CY (26 ТП)** — небольшие объемы продаж и невысокая стабильность. Запасы этой группы нужно создавать по остаточному принципу, т. е. после того, как выделены деньги на закупки товаров из более выгодных групп. Даже если в какой-то момент товара из этой группы не окажется на складе, потери компании будут невелики.
* **Группа CZ (1 677 ТП)** — самые низкие объемы выручки, при этом спрос невозможно спрогнозировать. Сюда относятся товары, которые уже перестали пользоваться спросом. Запасы по этой категории обычно не создают, работают под заказ. Но! В эту группу могут попасть и новинки. Они не пользуются спросом потому, что клиенты пока о них не знают. Эти товары еще не привезли в магазины, не выставили на витринах, не запустили маркетинговую кампанию и пр.



# **RFM-анализ**

**RFM-анализ** — метод анализа, позволяющий сегментировать клиентов по частоте и сумме покупок и выявлять тех, которые приносят больше денег.

Аббревиатура RFM расшифровывается:

* Recency — давность (как давно ваши клиенты что-то у вас покупали);
* Frequency — частота (как часто они у вас покупают);
* Monetary — деньги (общая сумма покупок).

По этим признакам можно разделить всех ваших клиентов на группы, понять, кто покупает у вас часто и много, кто — часто, но мало, а кто вообще давно ничего не покупал.

Как правило, небольшой процент пользователей реагирует на общие рекламные предложения. RFM-анализ и RFM-сегментация являются отличным методом прогнозирования реакции клиента и улучшения взаимодействия, а также повышение прибыли. RFM использует поведение покупателей, чтобы определить, как работать с каждой группой клиентов.

В нашем датафрейме 4 338 клиентов.

Определим давность покупки, частоту и общую сумму покупок для каждого клиента. Т.к. последняя дата покупки 09.12.2011, то считать давность начнем с 11.12.2021 для исключения отрицательных и нулевых значений количества дней с момента последней покупки. Далее поделим клиентов на 3 равные группы по 33 и 66 процентилю для каждого признака.

Далее сформируем сегменты по совокупности показателей R, F, M и в наш датафрейм добавим столбцы с данным группами:

**Давность заказа:**

* 1 — давние;
* 2 — «спящие» (относительно недавние);
* 3 — недавние.

**Частота покупок:**

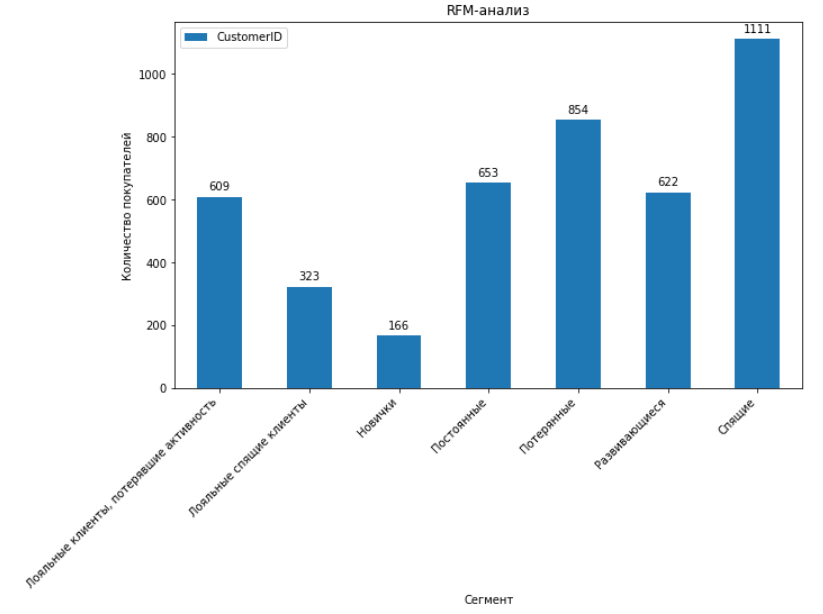
* 1 — разовые;
* 2 — редкие;
* 3 — частые.

**Сумма покупок:**

* 1 — низкий чек;
* 2 — средний чек;
* 3 — высокий чек.

Мы считали, что 1 — худшее значение, а 3 — лучшее. Получается три оценки по давности для каждого из трёх показателей. **Из совокупности трёх показателей с одинаковыми оценками складывается сегмент:**

* R1—F1—M(1–3) — потерянные;
* R1—F(2–3)—M(1–3) — лояльные клиенты, потерявшие активность;
* R2—F(1–2)—M(1–3) — спящие;
* R2—F(3)—M(1–3) — лояльные спящие клиенты;
* R3—F1—M(1–3) — новички;
* R3—F2—M(1–3) — развивающиеся;
* R3—F3—M3 — постоянные.



В результате RFM-анализа видим, что у нас **653 постоянных клиента**, которые совершают покупки часто и на большую сумму. Можно попросить оставить отзыв и сообщить о персональном обслуживании. Делать персональные скидки вовсе не обязательно.

Необходимо удержать интерес **622** **развивающихся клиентов.** Можно выяснить с помощью опроса, доволен ли клиент и какие у него пожелания. И другую информацию, которая пригодится для удержания: что ему интересно, какие у него потребности. Таким клиентам можно не предлагать персональные скидки, т.к. у них все шансы стать постоянными покупателями по полной стоимости. Лучше мотивировать чем-то другим. Например, шанс получить подарок в обмен на отзыв.

**Новичкам** **(166 клиентов)** можно предложить помощь в выборе продукта. Необходимо их перевести в ряды лояльных. Также можно поздравить с покупкой или поблагодарить за выбор нашей компании.

**Лояльным клиентам, потерявшим активность (609 клиентов),** **также лояльным спящим клиентам (323 клиента) и спящим** **(1 111 клиента)** можно предложить купон на скидку, бонусы, персональную товарную подборку или рассказать о распродаже

На **потерянных клиентов (854 клиента)** не стоит тратить много времени и усилий. Можно попробовать их вернуть, рассказав об акциях, скидках и распродажах или написать почему выгодно оставаться с нами.

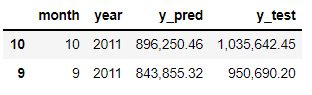
**Построение модели, предсказывающей объем продаж на будущие 3 месяца**

С помощью метода линейной регрессии, спрогнозируем объем продаж на будущие 3 месяца.

Подготовим датасет. Для этого сгруппируем объемы продаж по месяцам, оставим только необходимые столбцы InvoiceDate и TotalPrice и исключим данные за декабрь 2011, т.к. данные у нас не за полный месяц. Столбец Quantity в модели не используем, т.к. Quantity иTotalPrice имеют высокую корреляцию, что негативно может сказаться на нашей моделе.Далее добавим столбцы с номером месяца и годом. Разделим датасет на обучающую и тестовую выборки в соотношении 85/15.

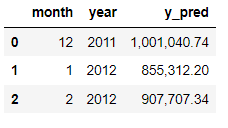
Обучим нашу модель и проверим качество модели с помощью метрики RMSE. **Среднеквадратическая ошибка (RMSE**) - квадратный корень из cредней квадратической ошибки MSE. Квадратный корень введен, чтобы масштаб ошибок был таким же, как масштаб целей.

Предскажем результат для тестовой выборки:



Рассчитаем RMSE: RMSE = 124 184.98215877629

Построим прогноз на будущие 3 месяца: декабрь 2011, январь 2012 и февраль 2012:



Модель предсказала следующий объем продаж: на декабрь 2011 – 1 000 040.74, на январь 2012 - 855 312.20 и февраль 2012 – 907 707.34. Модель может ошибиться в спрогнозированном значении на 124 184.98

# **Вывод**

В результате проведенных исследования можно дать следующие рекомендации для Департамента продаж и маркетинга:

1. Развитие продаж в других странах, т.к. объемы продаж за рубежом небольшие. Но предварительно оценить эффективность развития данного направления.
2. Оптимизировать товарный запас в соответствии с результатом ABC-XYZ-анализа.
3. Провести маркетинговую акцию для клиентов-новичков, лояльных клиентов, потерявшим активность, лояльных спящих клиентов и спящих. Таким клиентам необходимо напомнить о нашей Компании и сделать им выгодное предложение.

Построена модель прогнозирования продаж на будущие 3 месяца. По факту накопления данных по продажам за больший период времени, данную модель можно переобучить для более точного прогноза.

Также на основе очищенных данных подготовлен [**Дашборд по продажам E-commerce**](https://datalens.yandex/lllqulhhxqqce)**:**

