Университет

нетология

(Общество с ограниченной ответственностью)

Финальная работа по курсу

«Метрики, гипотезы, точки роста»

(поток AIC-41)

Лектор курса: Дмитрий Орлов

Эксперт курса: Арсений Сова

Студент: Сергей Черняев, DAU-37

2023 год

1. Описание данных: что отражают, их качество и полнота, некорректности и аномалии.

По условию задачи датасет содержит информацию по продажам интернет-магазина размещённого в Великобритании. В исходном датасете содержится **54214** записей.

В качестве инструмента анализа использовался язык **Python** с набором библиотек в дистрибутиве пакета **«Andconda»**.

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 54214 entries, 0 to 54213
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   InvoiceNo    54214 non-null  object
 1   StockCode    54214 non-null  object
 2   Description  54080 non-null  string
 3   Quantity     54214 non-null  int64
 4   InvoiceDate  54214 non-null  datetime64[ns]
 5   UnitPrice    54214 non-null  float64
 6   CustomerID   40643 non-null  Int64
 7   Country      54214 non-null  string
dtypes: Int64(1), datetime64[ns](1), float64(1), int64(1), object(2), string(2)
memory usage: 3.4+ MB
```

Илл.1 Информация об исходном наборе данных.

В датасете имеется восемь полей:

InvoiceNo - номер заказа. Тип данных - текстовое поле, поскольку некоторые значения представлены с пометкой "C" перед числовым значением. Пропущеных значений данное поле не имеет. Позиции из одного заказа имеют идентичные номера. Уникальных значений номеров заказов InvoiceNo: 14877

StockCode - код товара. Тип данных - текстовое поле. Пропущеных значений данное поле не имеет. Уникальных значений кодов товаров StockCode: 3351

Description - описание товара. Тип данных - текстовое поле. Имеется 134 пропущенных значения. Возможно попытаться восстановить используя код товара. Уникальных описаний товаров - 3402. Описаний больше чем кодов товара, при дальнейшем анализе выяснилось что даже в 10 самых популярных товарах есть по два описания к одному коду товара.

Quantity - количество позиции товара в заказе. Тип данных - целое число. Пропусков не имеется.

InvoiceDate - дата и время заказа товара. Тип данных - datetime64. Пропусков не имеется. Самая раняя дата и время заказа: **01 декабря 2019 года в 08:26:00.**

Наиболее поздняя дата и время заказа: **09 декабря 2020 года в 12:50:00**. Таким образом в распоряжении аналитика имеется практически полный год данных для проведения анализа.

UnitPrice - цена за единицу товара. Пропусков не имеется. Тип данных - число с плавающей точкой. Если исключить из анализа технические записи возвратов, комиссий, ручных записей и т.д., то минимальная цена составляет 4 пенса за позицию: **POPART WOODEN PENCILS ASST**, а максимальная цена в 195 фунтов стерлингов за товар **LOVE SEAT ANTIQUE WHITE METAL.**

CustomerID - идентификатор клиента. Уникальных идентификаторов клиентов - 3820. Имеется 13571 пропуск, что составляет 25% от всего набора данных. Восстановление невозможно. Серьёзный недостаток, который повлияет на определение характеристик клиентов.

Country - страна-местонахождение клиента. Пропусков не наблюдается. Имеется 37 уникальных значений:

'United Kingdom' 'Australia' 'Netherlands' 'Norway' 'EIRE' 'Germany'
'France' 'Switzerland' 'Spain' 'Poland' 'Italy' 'Belgium' 'Lithuania'
'Japan' 'Portugal' 'Iceland' 'Channel Islands' 'Denmark' 'Cyprus'
'Sweden' 'Finland' 'Austria' 'Israel' 'Greece' 'Hong Kong' 'Singapore'
'Lebanon' 'United Arab Emirates' 'Czech Republic' 'Canada' 'Unspecified'
'Brazil' 'USA' 'European Community' 'Bahrain' 'Malta' 'RSA'

Первой очевидной аномалией можно считать **нулевую цену товара**. Возможно таким образом осуществлялась коррекция остатков, но это не записи о реальных продажах или возвратах.

Решение: 225 записей с нулевой ценой товара исключены из дальнейшего анализа.

В процессе анализа кодов и цены товаров определилась группа "технических" записей связанных с пересылкой, ручным изменением данных, накладными расходами, дисконтами, оплатой маркетплейса Amazon и т.д.. Данные операции записаны под кодами товаров 'DOT','M', 'CRUK', 'BANK CHARGES','D', 'POST', 'AMAZONFEE'. Поскольку эти записи не отражают продажу реальных товаров их нужно исключить из датасета.

Решение: 272 записи техническими проводками исключены из дальнейшего анализа.

Отрицательные значения количества товаров не являются ошибочными, поскольку вместе с префиксом «**C**» в столбце **InvoiceNo** они указывают на возврат товара.

2. Анализ данных. Определение ключевых показателей

Для дальнейшего анализа вычислим стоимость позиции в заказе и запишем её в отдельное поле **Cost**. Выделим отдельно дату заказа в поле **InvoiceData:**

Общие показатели за имеющийся период:

1. Сумма продаж, количество единиц товаров в заказах, количество всех строк позиций клиентов по (топу 10 стран):

	Cost_sum	Quantity_sum	InvoiceNo_count
Country			
United Kingdom	869053.46	455218	48443
EIRE	27321.44	14304	757
Netherlands	26399.18	20924	229
Germany	20864.82	12102	875
France	18262.48	11344	835
Australia	17156.59	10009	123
Switzerland	5695.90	3528	184
Spain	5551.22	2921	246
Sweden	4816.52	4437	46
Belgium	3628.66	2371	196

Табл. 1. Сумма продаж, количество единиц товаров в заказах, количество всех строк позиций клиентов по 10 лидирующим странам.

Доля Великобритании в общем объёме продаж составляет 87.0%. Доля Великобритании в общем количестве позиций заказов - 93.3%. Значительный разрыв между Великобританией Ирландией и остальными странами объясняется размещением интернет-магазина в Великобритании и его региональной ориентацией. Аномальной представляется ситуация по Гонконгу - нет определяемых идентификаторов клиентов, и в США - отрицательное количество заказанных единиц товаров.

Незаполненное поле **CustomerID** для заказов из Гонконга вероятно свидетельствует о том, что все заказы оформлялись без регистрации в Интернет-магазине. Можно предположить, что и в других подобных случаях **отсутствие CustomerID говорит о том, что покупка осуществлялась без регистрации на сайте магазина.**

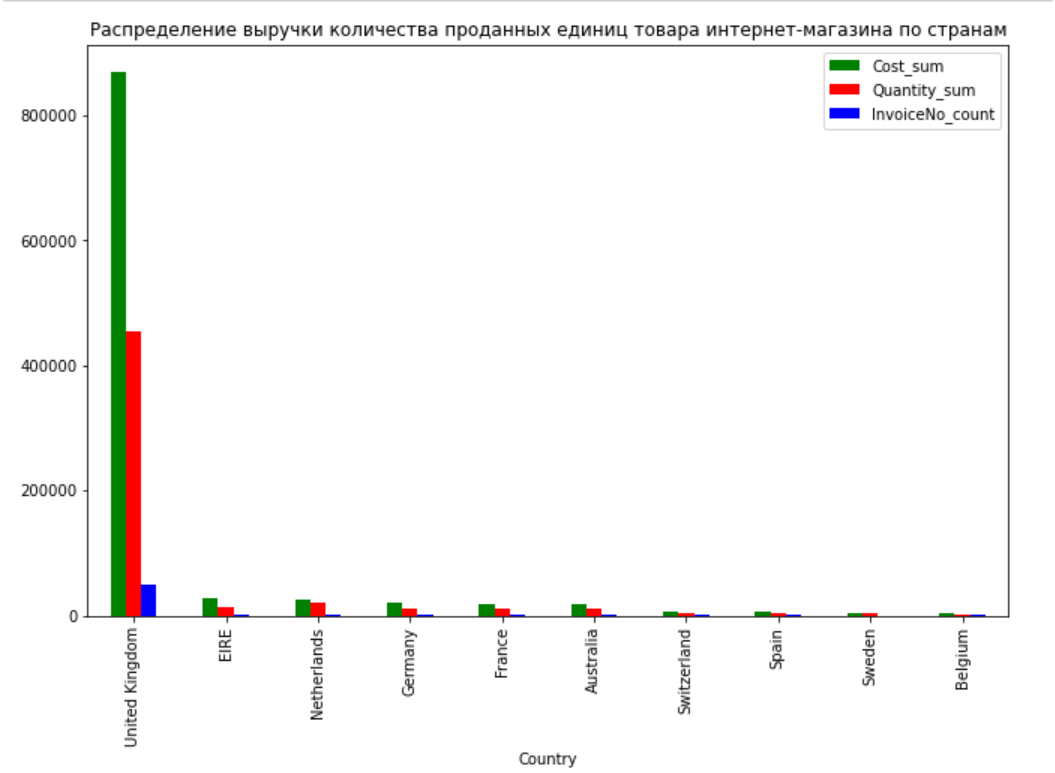
Процент возврата по всем продажам составляет 2,0%.

Помесячный анализ продаж показывает, что объем выручки возрастает к концу года. Из 20 лучших дней по продажам 5 дней приходится на ноябрь 4 дня на декабрь и по два дня на сентябрь и октябрь.

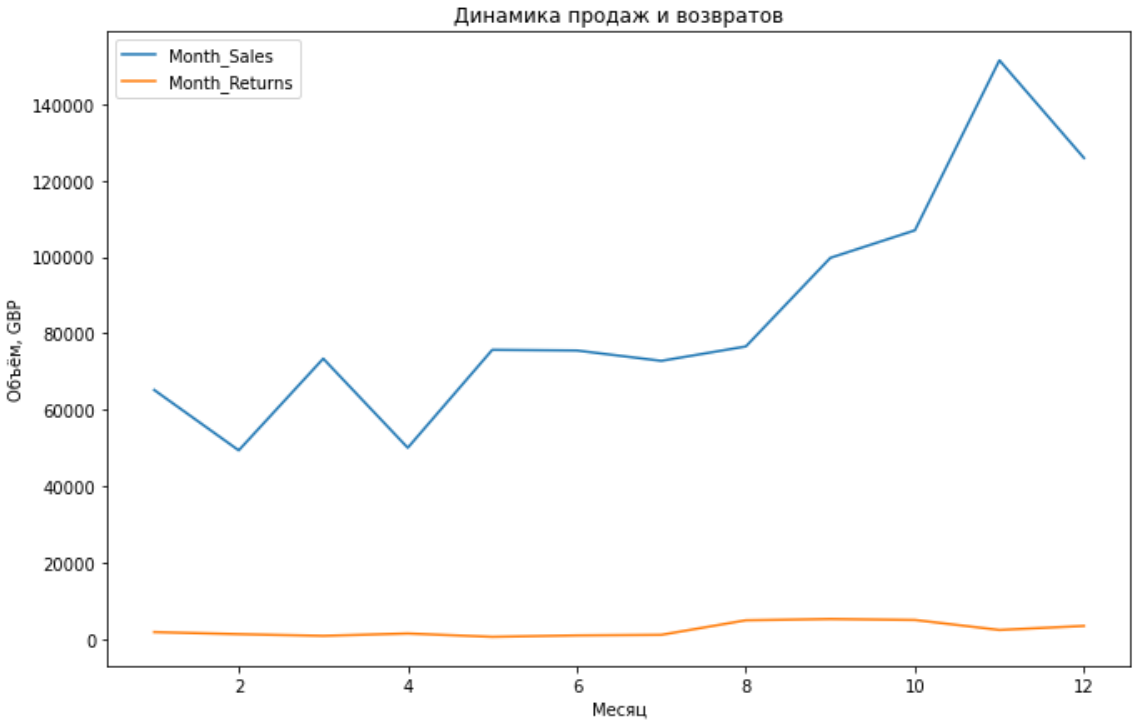
```
InvoiceData
2020-11-14    12028.34
2020-11-07    11930.73
2020-12-07    10606.61
2020-07-28     9708.64
2020-12-08     9592.91
2020-10-05     9449.37
2020-12-05     8920.57
2020-06-30     8588.98
2020-08-17     8530.31
2019-12-07     8475.61
2020-11-09     7901.87
2020-09-22     7763.77
2020-09-20     7444.12
2020-03-29     7322.38
```

```
                2020-10-20     7311.63
                2020-01-14     7224.09
                2020-11-23     7128.76
                2020-05-24     6843.89
                2020-05-12     6772.05
2020-11-03     6753.98
Name: Cost, dtype: float64
```

Табл. 2. Лучшие 20 дней по объму выручки.



Илл. 2 Распределение выручки по странам.



Илл. 2 Помесячная динамика продаж и возвратов по всем странам.

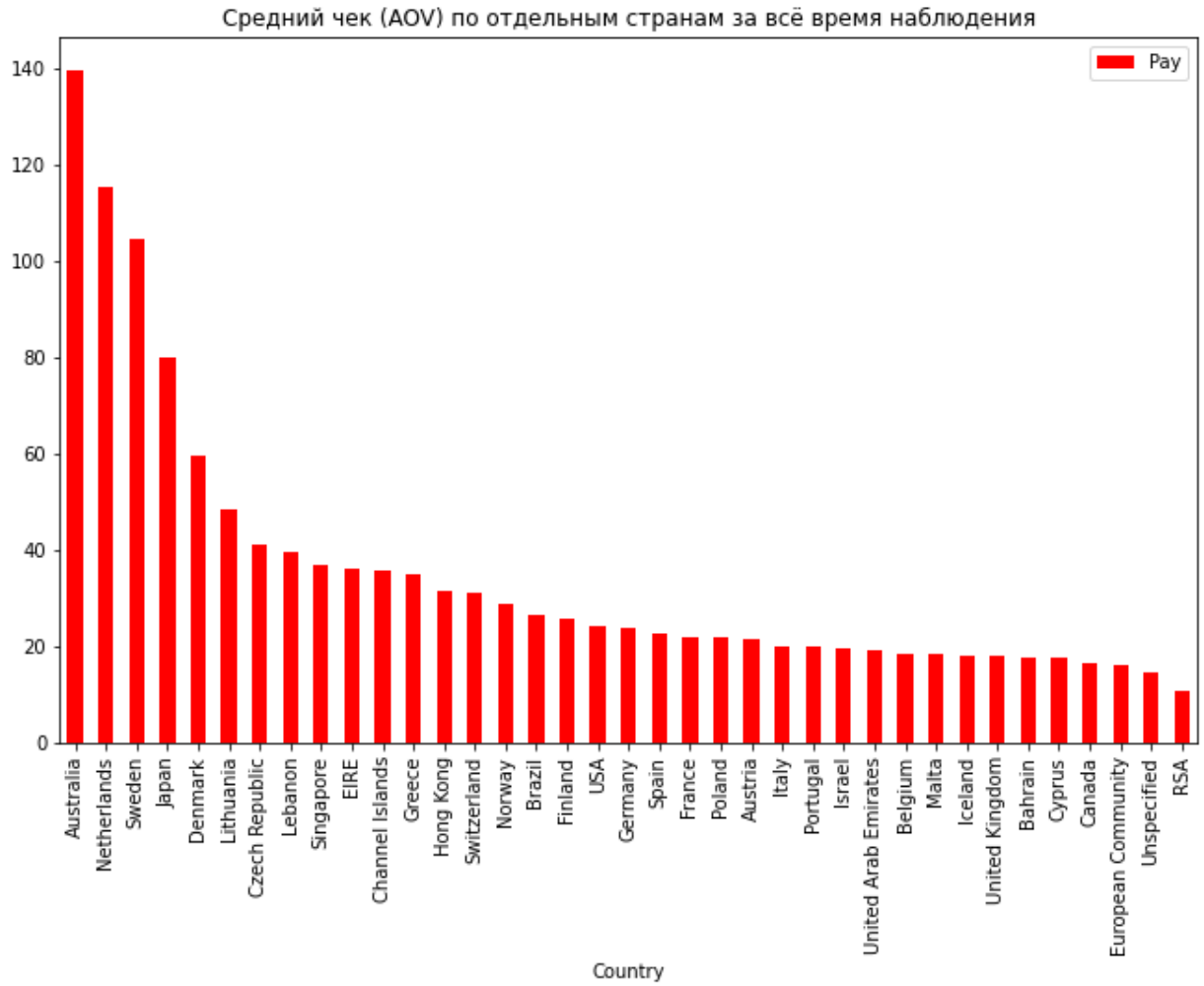
Продажи возрастают к окончанию года в связи с тем, что начинаются период подготовки к рождественским и новогодним праздникам. Поскольку полноценной статистики за декабрь-месяц не имеется, достоверно определить уровень падения к концу месяца не представляется возможным. Начало года показывает спад в уровне продаж. Уровень возвратов связан с уровнем продаж. С ростом продаж пропорционально возрастает и уровень возвратов в среднем он составляет **1,7%. Коэффициент корреляции на уровне 0,45 по шкале Чеддока говорит об умеренной положительной связи между продажами и возвратами.**

Средний чек **(AOV - average order value)** по всем странам за весть период наблюдения держится на уровне **19,36 фунтов стерлингов.** Высокий относительно общего уровня показатель AOV по Австралии, Нидерландам, Японии, Дании и Литве может быть объяснён тем, что доставка в данные регионы занимает длительное время и клиенты оформляют заказы на эксклюзивные товары, которые недоступны в этих странах.

Country	AOV
Australia	139.484472
Netherlands	115.280262
Sweden	104.706957
Japan	80.154815
Denmark	59.671765
Lithuania	48.600000
Czech Republic	41.100000
Lebanon	39.600000

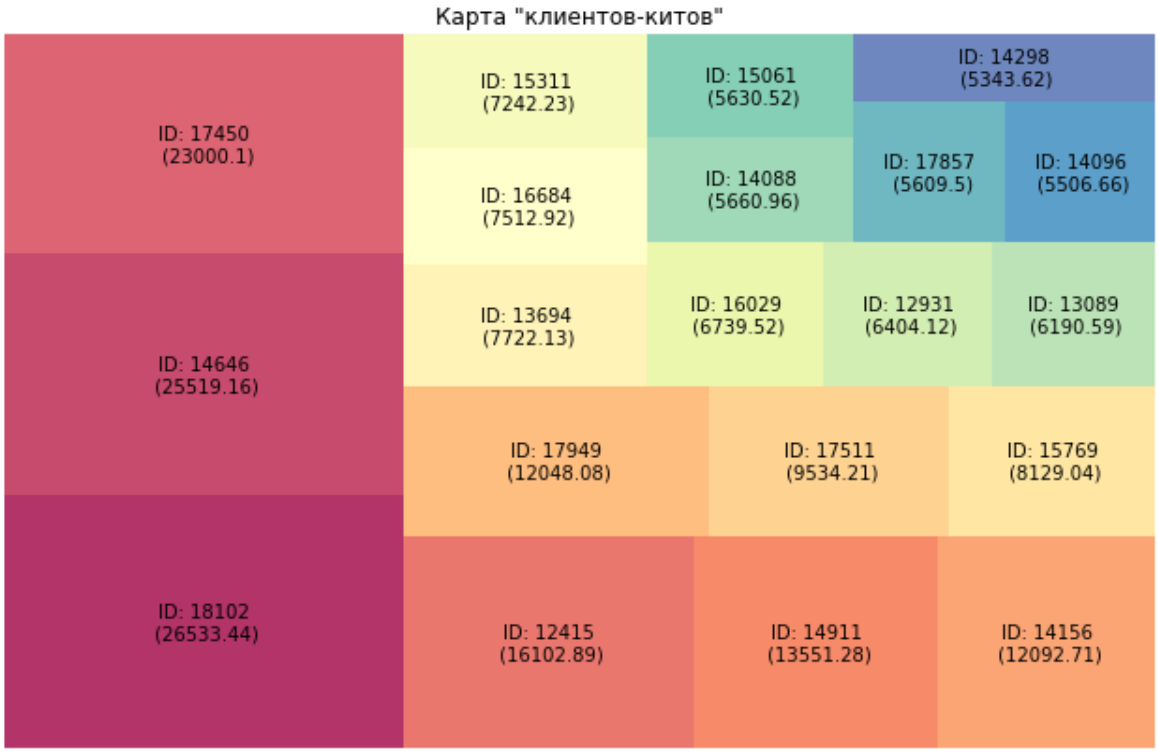
Country	AOV
Singapore	36.795000
EIRE	36.091731
Channel Islands	35.530247
Greece	35.058571
Hong Kong	31.393214
Switzerland	30.955978
Norway	28.640404
Brazil	26.600000
Finland	25.722857
USA	24.245000
Germany	23.845509
Spain	22.565935
France	21.871234
Poland	21.704524
Austria	21.588958
Italy	19.938784
Portugal	19.840207
Israel	19.693333
United Arab Emirates	19.284000
Belgium	18.513571
Malta	18.271429
Iceland	18.119333
United Kingdom	17.939712
Bahrain	17.700000
Cyprus	17.632500
Canada	16.472308
European Community	16.180000
Unspecified	14.450556
RSA	10.535000

Табл. 3 Средний (AOV) чек по отдельным странам за всё время наблюдения.

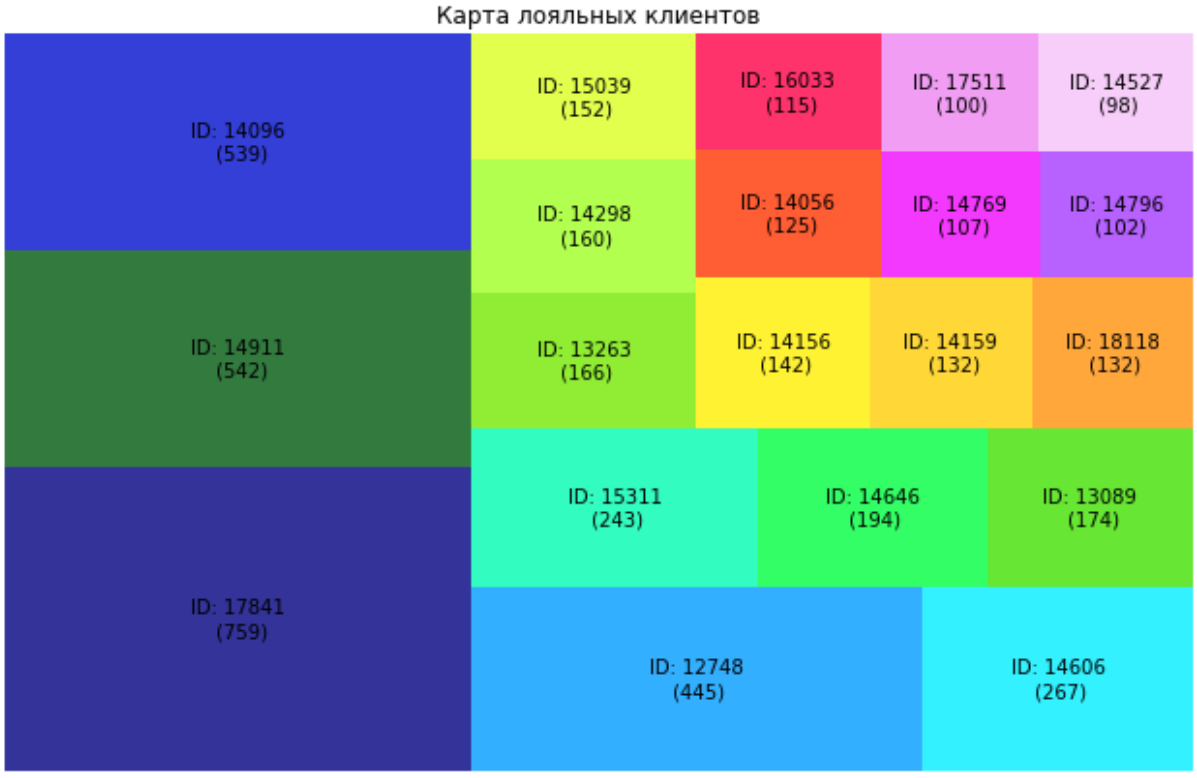


Илл. 3 Средний (AOV) чек по отдельным странам за всё время наблюдения.

Анализ активности покупателей показывает, что среди клиентов сформировано ядро потребителей дающих как существенную прибыль максимально до **26533,44 фунтов стерлингов**, так и заказывающих за год **до 759 позиций.**



Илл. 4 Клиенты-киты (топ 20) за всё время наблюдения.



Илл. 5 Лояльные клиенты (топ 20) за всё время наблюдения.

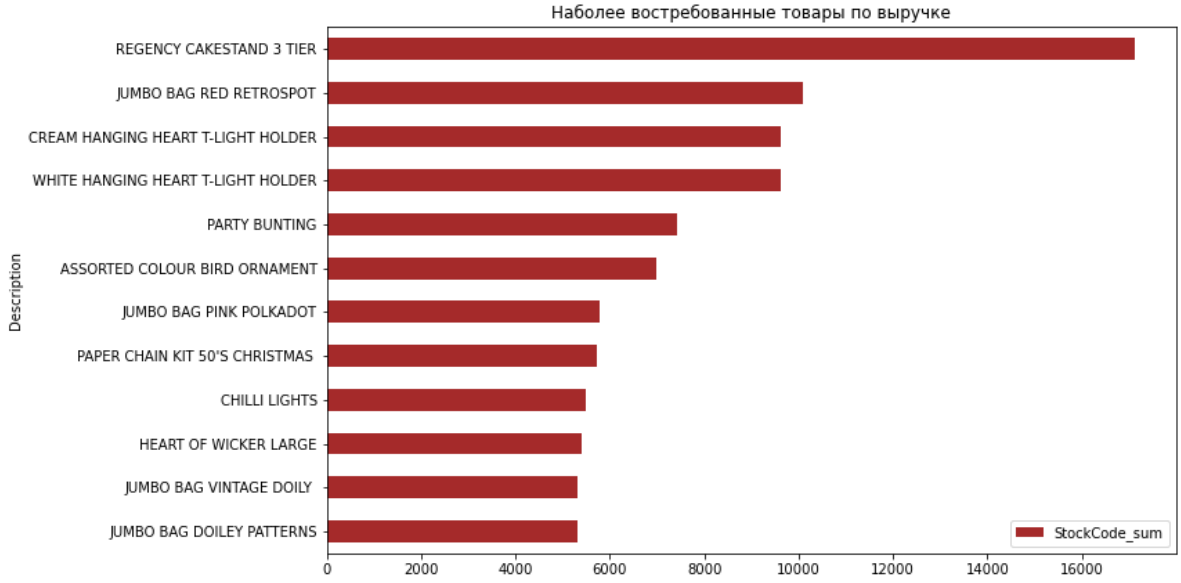
Частота повторных покупок (**RPR — repeat purchase rate**) составляет 27,3%

Наиболее востребованные товары по выручке (топ 10) представлены в табл. 4

StockCode	StockCode_sum	Description
22423	17135.000000	REGENCY CAKESTAND 3 TIER
85099B	10092.030000	JUMBO BAG RED RETROSPOT
85123A	**9609.590000**	**WHITE HANGING HEART T-LIGHT HOLDER**
85123A	**9609.590000**	**CREAM HANGING HEART T-LIGHT HOLDER**
47566	7412.360000	PARTY BUNTING
84879	6975.590000	ASSORTED COLOUR BIRD ORNAMENT
22386	5794.930000	JUMBO BAG PINK POLKADOT
22086	5714.650000	PAPER CHAIN KIT 50'S CHRISTMAS
79321	5501.530000	CHILLI LIGHTS
22470	5393.930000	HEART OF WICKER LARGE
23203	5312.760000	JUMBO BAG DOILEY PATTERNS
23203	5312.760000	JUMBO BAG VINTAGE DOILY

Табл. 4 Наиболее востребованные товары по выручке (топ 10).

В выборке выявлена некорректность, связанная с тем, что уникальному коду товара могут соответствовать несколько разных описаний. Так например для кода 85123A, имеется два описания: первое - **WHITE HANGING HEART T-LIGHT HOLDER, второе - CREAM HANGING HEART T-LIGHT HOLDER**.



Илл. 6 Наиболее востребованные товары по выручке.

3. Описание потенциального бизнес-заказчика, его потребностей и проблем и целей развития бизнеса.

Потенциальным бизнес-заказчиком в данном кейсе может выступать как владелец бизнеса и операционный менеджмент, так и потенциальный инвестор – венчурный фонд, бизнес-ангел и т.д. Данный анализ не предоставляет высокой степени детализации, но даёт понимание общих проблем и «узких мест» бизнеса.

Потенциальные проблемы выявленные при анализе:

<u>В сфере продаж и маркетинга:</u>

- крайне неравномерный охват клиентской базы, 87,0% выручки и 93, 3% позиций заказов поступает из страны местонахождения;
- слабо представлены близлежащие страны Евросоюза;
- практически не охвачены американский и азиатский рынки;
- сезонность продаж со спадами в феврале и апреле.

<u>В сфере культуры сбора данных:</u>

- неудачная концепция записи возвратов через префикс в номере заказа вместо отдельного поля;
- сдвоенные описания товаров с совпадающим артикулом;
- наличие в данных технических транзакций на значительные суммы по банковскими операциям, доставке, дисконту не связанными с осуществлением продаж товаров
- пропуск идентификаторов клиентов, предположительно при покупке без регистрации, что вносит неопределённость в анализ клиентской базы.

<u>Цели развития бизнеса:</u>

- расширение географии продаж с увеличением выручки в европейских странах и последующим выходом на американский и азиатские рынки;
- борьба с сезонностью;
- увеличение показателя удержания клиентов (CRR).

4. Описание заинтересованных лиц отчёта с обоснованием выбранных метрик (системы метрик).

Заинтересованными в отчете сторонами могут быть: отделы продаж, маркетинга, руководитель.

При дальнейшем развитии аналитики необходимо дополнить систему метрик показателями описывающими процесс привлечения аудитории через технологии интернет-маркетинга:

- **CPC (Cost Per Click)** – стоимость клика пользователя по объявлению;
- **CTR (Click-Through Rate)** – соотношение числа кликов к числу показов;

- **CR (Conversion Rate)** – отношение пользователей, которые совершили целевое действие, к общему количеству посетителей сайта;
- **CAC (Customer Acquisition Cost)** – сумма, которую компания потратила на маркетинг, чтобы привлечь покупателя;
- **CPL (Cost Per Lead)** – стоимость потенциального клиента, который интересуется товаром;
- **ROI (Return on Investment)** – интегральный показатель, отражающий возврат инвестиций в маркетинговые активности. Метрика показывает, заработала ли компания на продажах больше денег, чем вложила в продвижение продукта.

Обязательные дополнительные метрики помимо рассчитанных **AOV** и **RPR** которые характеризуют клиентскую базу интернет-магазина:

- **ARPU (Average Revenue Per User)** – средний доход от одного клиента;
- **Lifetime** – период, в течение которого покупатель осуществляет заказы;
- **LTV (Lifetime Value)** – пожизненная ценность клиента;
- **CRR (customer retention rate)** – показатель удержания клиентов;
- **CAR (Cart Abandonment Rate)** – процент пользователей, которые добавили товары в корзину и не купили их.

С целью получения исходной информации по расчёту добавочных метрик необходимо продумать возможность подключения внешней аналитики сайта на основе «Google Analytics» и внедрения BI-инструментария с доступной и понятной визуализацией позволяющей получить необходимый уровень детализации для разных категорий пользователей.