Анализ товарного ассортимента

Кирилл Жбаков

2022

Цели и задачи

Цель – предоставление рекомендаций покупателям по товарным предложениям.

Задачи:

- обзор данных;
- предобработка данных;
- исследовательский анализ данных;
- статистические гипотезы;
- составление презентации.

Данные

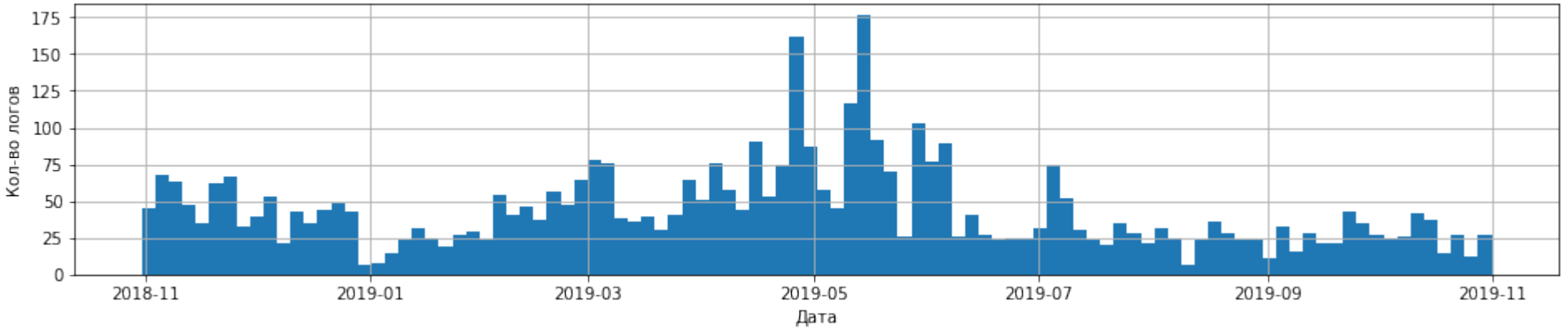
	date	customer_id	order_id	product	quantity	price
0	2018100100	ee47d746-6d2f-4d3c-9622-c31412542920	68477	Комнатное растение в горшке Алое Вера, d12, h30	1	142.0
1	2018100100	ee47d746-6d2f-4d3c-9622-c31412542920	68477	Комнатное растение в горшке Кофе Арабика, d12,...	1	194.0
2	2018100100	ee47d746-6d2f-4d3c-9622-c31412542920	68477	Радермахера d-12 см h-20 см	1	112.0
3	2018100100	ee47d746-6d2f-4d3c-9622-c31412542920	68477	Хризолидокарпус Лутесценс d-9 см	1	179.0
4	2018100100	ee47d746-6d2f-4d3c-9622-c31412542920	68477	Циперус Зумула d-12 см h-25 см	1	112.0

- *date* — дата заказа;
- *customer_id* — идентификатор покупателя;
- *order_id* — идентификатор заказа;
- *product* — наименование товара;
- *quantity* — количество товара в заказе;
- *price* — цена товара.

```
----------Тип данных----------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6737 entries, 0 to 6736
Data columns (total 6 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   date         6737 non-null   int64  
 1   customer_id  6737 non-null   object 
 2   order_id     6737 non-null   int64  
 3   product      6737 non-null   object 
 4   quantity     6737 non-null   int64  
 5   price        6737 non-null   float64
dtypes: float64(1), int64(3), object(2)
memory usage: 315.9+ KB

None

----------Пропуски в данных----------
Пропусков нет

----------Количество явных дубликатов----------
0
```

Исследовательский анализ данных

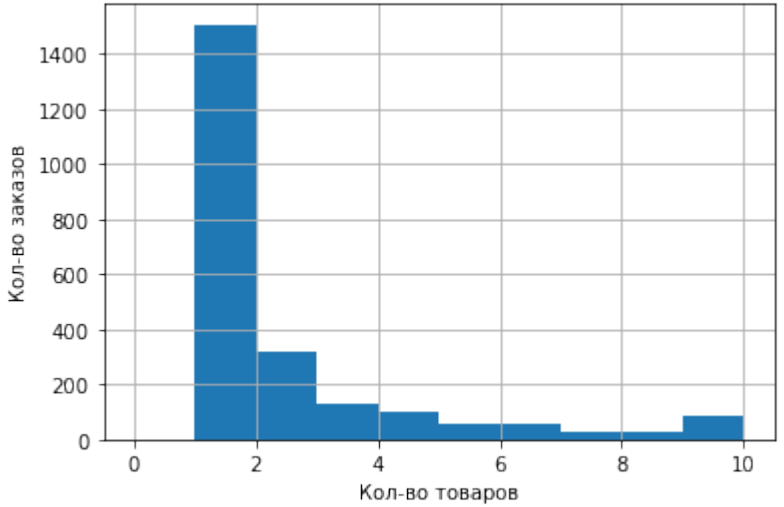
Анализ полноты данных



- Данные собраны с 2018-10-01 00:00:00 по 2019-10-31 16:00:00.
- Данные полные. В мае и июне наблюдается повышенное количество логов.
- Исключены данные октября 2018 года для дальнейшего анализа.
- Обработаны аномальные значения.

Основные характеристики датасета

- Всего 2279 покупателей/я .
- Всего 2552 заказа/ов.
- Среднее количество заказов на покупателя: 1.12.
- Среднее количество товаров в заказе 5.80.
- Количество заказов на покупателя распределено в диапазоне от одного до 106. При этом основная масса покупателей сделала заказ только один заказ.
- Количество товаров в заказе распределено в диапазоне от одного до 1000. У основной массы покупателей имеется лишь один товар в заказе.



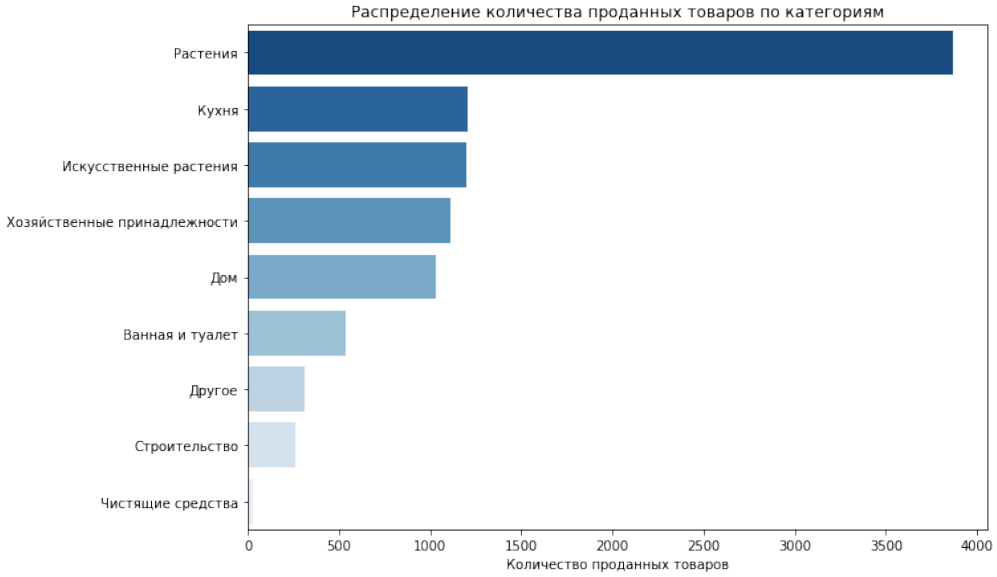
Категоризация продукции

Выделены следующие категории:

- 'Искусственные растения'
- 'Растения'
- 'Хозяйственные принадлежности' - товары для активных действий по дому (щётки, тазики, сумки).
- 'Ванная и туалет'
- 'Чистящие средства'
- 'Дом' - вещи для обеспечения уюта в доме, как правило статичные (постельное бельё, мебель, декор).
- 'Кухня'
- 'Строительство'
- 'Другое' (те товары, которые не вошли в категорию выше)

Количество логов по категориям

category	
Растения	3838
Кухня	1208
Искусственные растения	1196
Хозяйственные принадлежности	1109
Дом	1030
Ванная и туалет	538
Другое	309
Строительство	262
Чистящие средства	26



Анализ выручки по категориям

Суммарная выручка

category	
Хозяйственные принадлежности	₽ 1,198,079
Растения	₽ 523,137
Дом	₽ 429,115
Кухня	₽ 353,946
Ванная и туалет	₽ 341,855
Строительство	₽ 225,178
Другое	₽ 81,361
Искусственные растения	₽ 63,920
Чистящие средства	₽ 3,643

Name: revenue, dtype: object

Средняя выручка

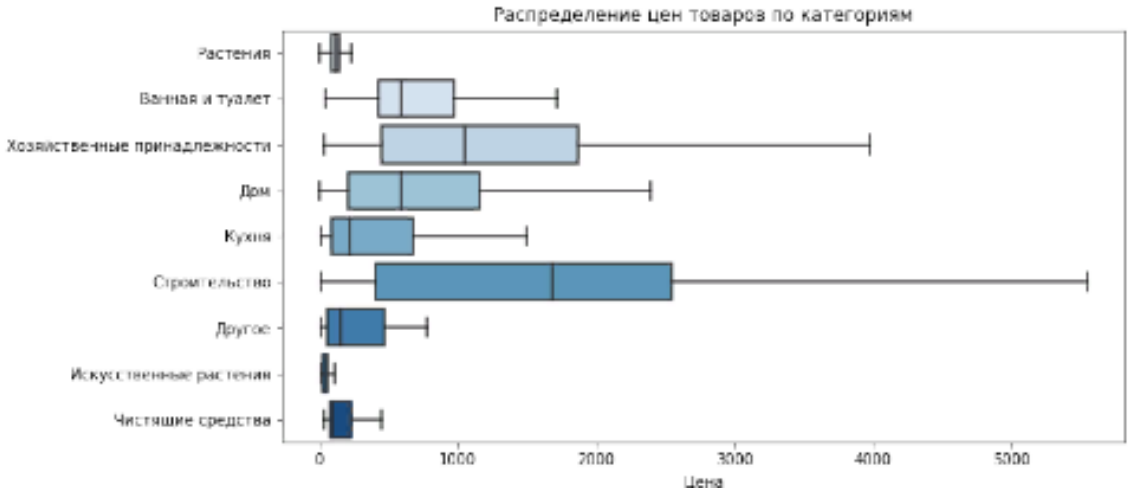
category	
Строительство	₽ 2,680.69
Хозяйственные принадлежности	₽ 2,058.55
Дом	₽ 1,430.38
Ванная и туалет	₽ 1,203.71
Кухня	₽ 1,053.41
Другое	₽ 774.87
Искусственные растения	₽ 318.01
Чистящие средства	₽ 260.21
Растения	₽ 244.57

Name: revenue, dtype: object

Средняя цена товара по категории

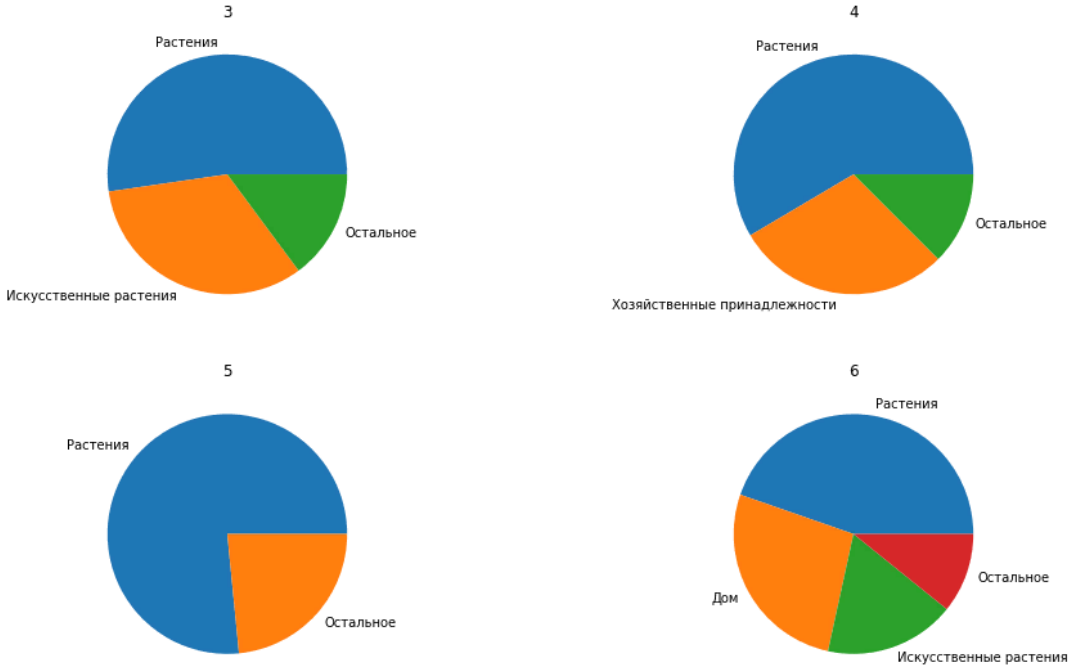
category	
Строительство	₽ 1,740.04
Хозяйственные принадлежности	₽ 1,485.04
Ванная и туалет	₽ 946.64
Дом	₽ 805.32
Кухня	₽ 532.32
Другое	₽ 506.33
Чистящие средства	₽ 178.21
Растения	₽ 147.42
Искусственные растения	₽ 72.27

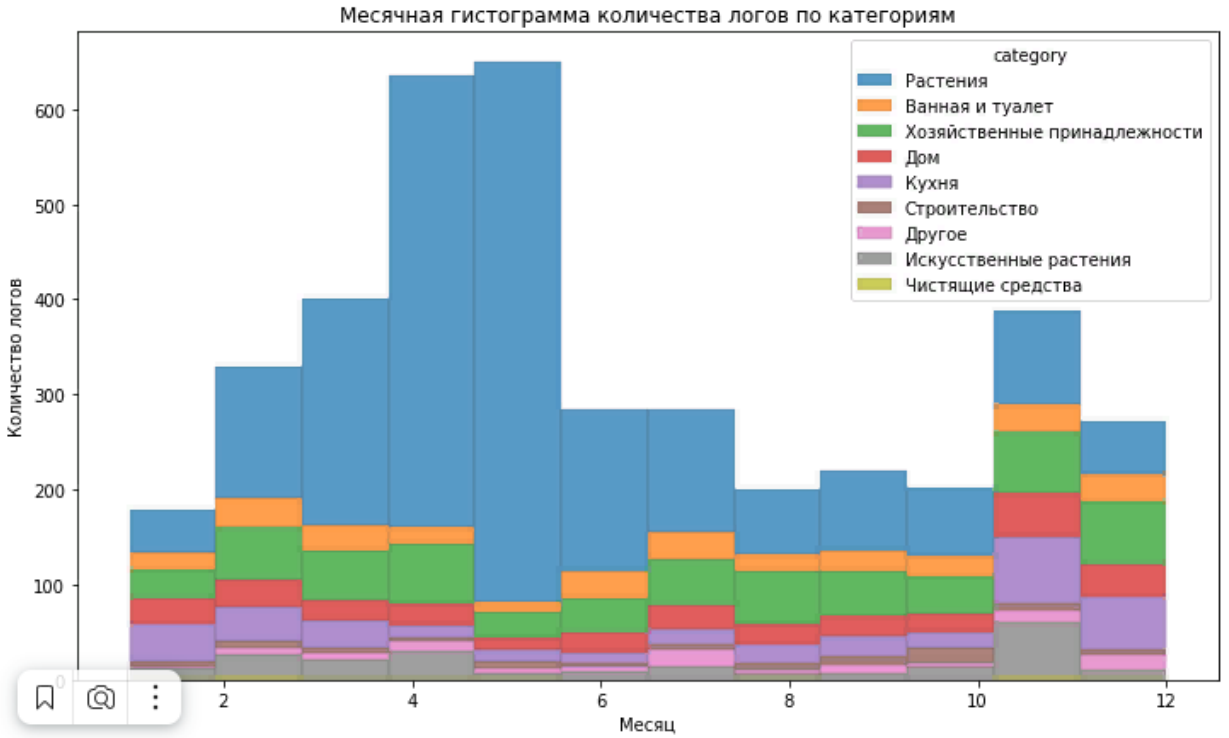
Name: price, dtype: object



Анализ месячной сезонности

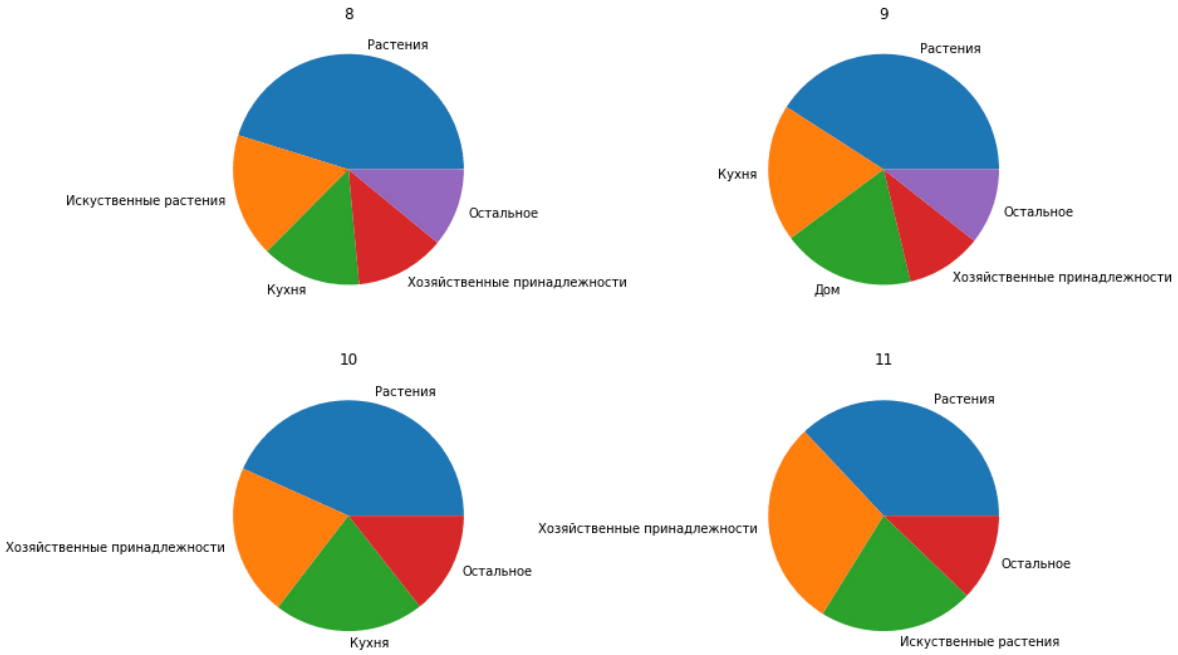
- Наблюдается повышенное количество покупок в ноябре и декабре (это конец 2018 года).
- Популярность растений выше обычного с марта по май - особенно в апреле и мае. Скорее всего интерес проявляют дачники, покупая рассаду.
- Заинтересованность искусственными растениями проявляется в октябре и ноябре, что может быть связано с подготовкой к новому году.
- В декабре и январе немного увеличиваются продажи кухонных, хозяйственных принадлежностей и товаров категории "Дом" что может быть связано с новогодними праздниками. Домашняя утварь - распространённый подарок.
- В остальных категориях месячная сезонность не проявляется.

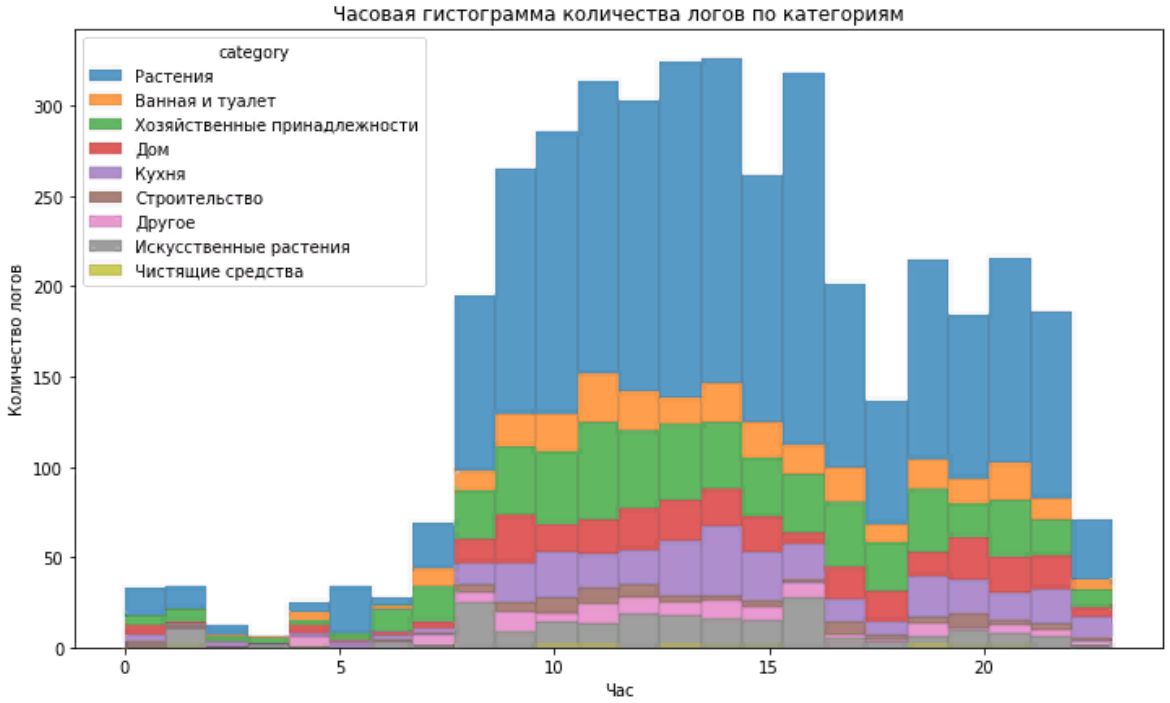




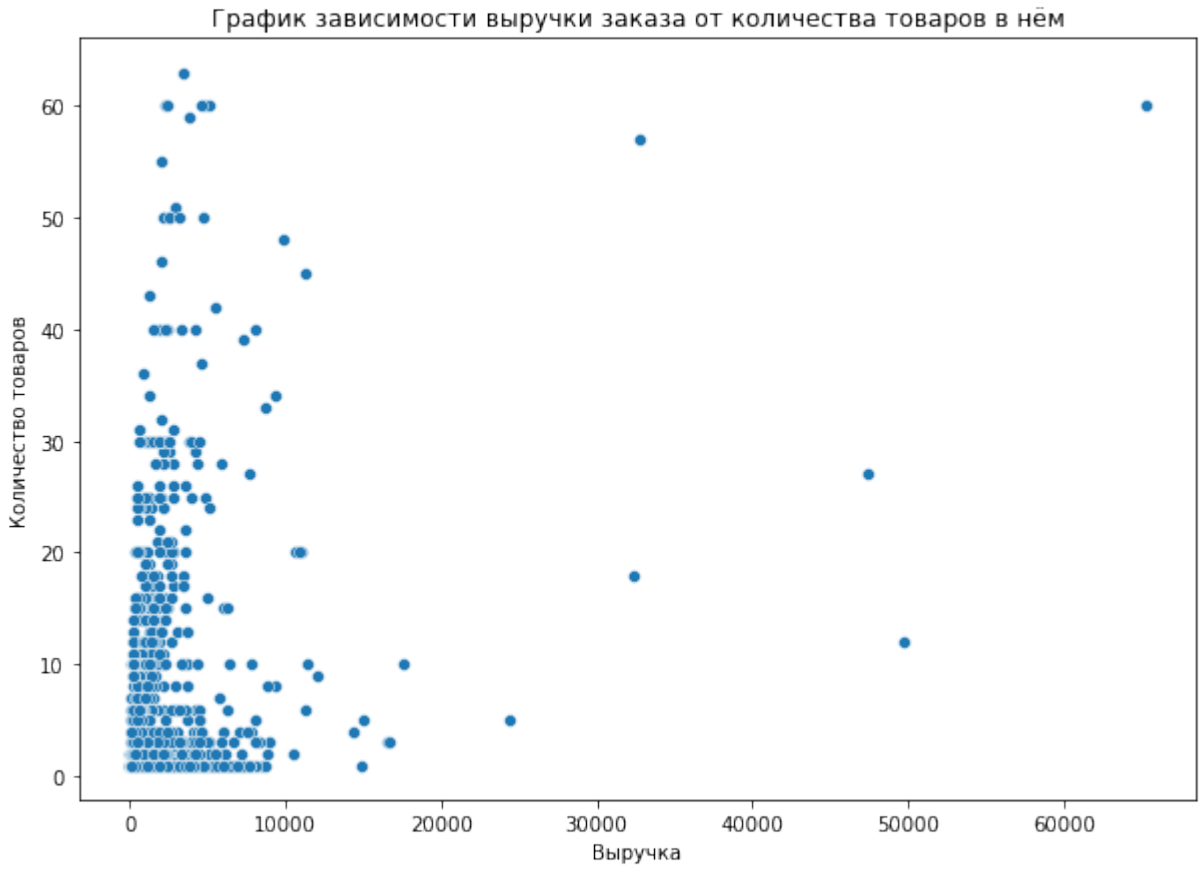
Анализ часовой сезонности

- активность покупателей возрастает с 7 утра и достигает в 11 часов утра "плато", которое длится до 17 часов;
- наблюдается заметный провал активности в 18 часов, и это может быть связано с тем, что потребители находятся на пути с работы домой;
- после 18 00 клиенты проявляют меньшую активность - дома нужно отдыхать=);
- в утренние часы, кроме растений популярны хоз. принадлежности;
- часовая сезонность в остальных категориях не обнаружилась.





Зависимость выручки заказа от количества товаров



- Чёткой зависимости между количеством товаров и выручкой заказа не наблюдается.
- Однако, средняя выручка заказа с несколькими товарами больше.
- Среднее группы с одним заказом: ₽ 1057.
- Среднее группы с заказами > 1: ₽ 1783.

Проверка статистических гипотез

Проверка сезонности доли товаров категории "Растения"

H0 - доля купленных товаров категории Растения с марта по май **>** доли купленных товаров с ноября по февраль и с июня по октябрь в объединении.
H1 - доля купленных товаров категории Растения с марта по май **<=** доли купленных товаров с ноября по февраль и с июня по октябрь в объединении.

- Доля категории "Растения" первой группы: 63%
- Доля категории "Растения" второй группы: 27%

Результат:
- p-значение: 0.0
- Не отвергаем нулевую гипотезу
- Доля купленных товаров категории Растения с марта по май > доли купленных товаров с ноября по февраль и с июня по октябрь в объединении.

Проверка сезонности средней выручки товаров в категории "Растения"

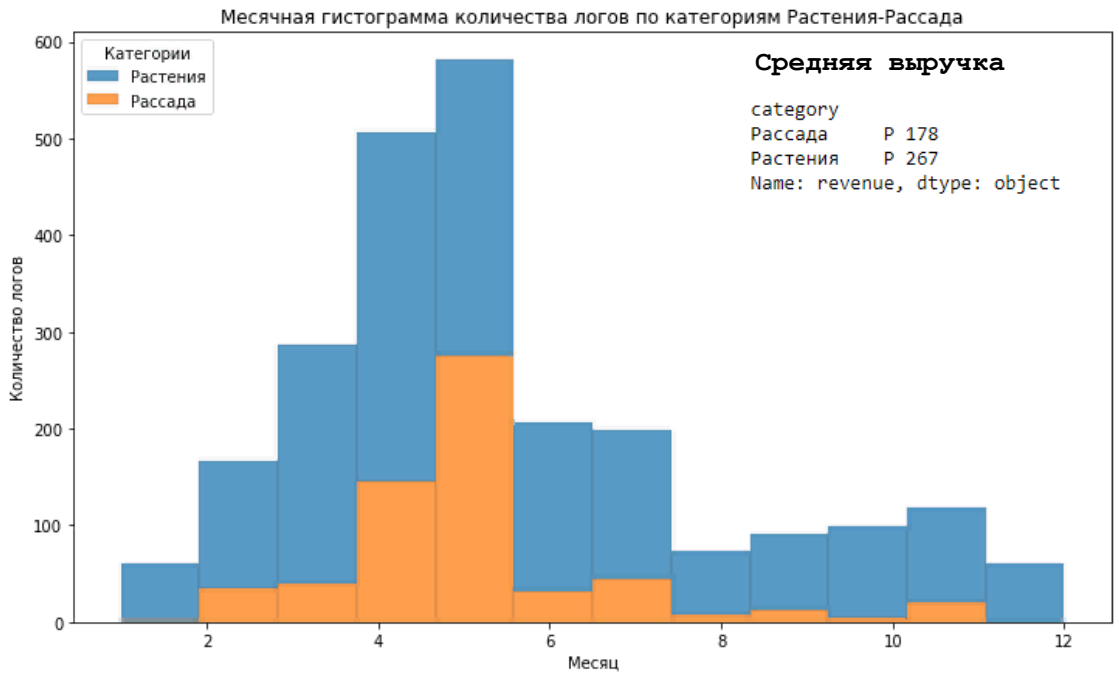
H0 - средняя выручка купленных товаров категории Растения с марта по май (первая группа) **>** средней выручки купленных товаров с ноября по февраль и с июня по октябрь в объединении (вторая группа).

H1 - средняя выручка купленных товаров категории Растения с марта по май (первая группа) **<=** средней выручки купленных товаров с ноября по февраль и с июня по октябрь в объединении (вторая группа).

- Среднее первой группы: 207
- Среднее второй группы: 300
- Дисперсия первой группы: 137,592
- Дисперсия второй группы: 625,035

Результат:

- p-значение: 0.0006503105574927675
- Отвергаем нулевую гипотезу
- Средняя выручка купленных товаров категории Растения с марта по июнь (первая группа) <=
средней выручки купленных товаров с ноября по февраль и с июня по октябрь в объединении (вторая группа).



Заказы с одним товаром и несколькими

H0 - средняя выручка заказов с одним товаром **>** средней выручки заказов, в которых товаров больше одного.
H1 - средняя выручка заказов с одним товаром **<** средней выручки заказов, в которых товаров больше одного.
H2 - средняя выручка заказов с одним товаром **=** средней выручки заказов, в которых товаров больше одного.

- Среднее группы с одним заказом: 1057
- Среднее группы с заказами > 1: 1788
- Дисперсия группы с одним заказом: 1,871,778
- Дисперсия группы с заказами > 1: 15,451,230

Результат:
- p-значение: 2.3408324388386094e-08
- Отвергаем нулевую и вторую гипотезы
- Средняя выручка заказов с одним товаром < средней выручки заказов, в которых товаров больше одного.

Выводы

Обзор данных:

- все столбцы присутствуют;
- необходимо обработать формат даты;
- в таблице 6737 логов.

Предобработка данных:

- изменён тип данных столбца с датами;
- добавлены столбцы с месяцами и часами с целью дальнейшего анализа сезонности.

Исследовательский анализ данных:

анализ полноты данных:

- данные собраны с 2018-10-01 00:00:00 по 2019-10-31 16:00:00;
- данные полные;
- в мае и июне наблюдается две волны всплеска количества логов;
- удалены логи за октябрь 2018.

исключение выдающихся значений:

- Суммарно было удалено 7.6% логов. Было принято, что удалённые покупатели это оптовики. Таким образом, для анализа остались только обычные потребители.

категоризация товаров:

- 'Искусственные растения'
- 'Растения'
- 'Хозяйственные принадлежности' - товары для активных действий по дому (щётки, тазики, сумки).
- 'Ванная и туалет'
- 'Чистящие средства'
- 'Дом' - вещи для обеспечения уюта в доме, как правило статичные (постельное бельё, мебель, декор).
- 'Кухня'
- 'Строительство'
- 'Другое' (те товары, которые не вошли в категорию выше)

изучение распределения количества товаров по заказам и покупателям:

- всего 2279 покупателей/я;
- всего 2552 заказа/ов;
- среднее количество заказов на покупателя: 1.12;
- среднее количество товаров в заказе 5.80;
- наиболее популярными категориями оказались "Растения", "Кухня", "Искусственные растения".

анализ выручки каждой категории:

- несмотря на популярность растений наибольшую выручку принесли хозяйственные товары;
- наибольшая средняя выручка характерна для категории "Строительство" и "Хозяйственные принадлежности";
- наиболее дорогие товары характерны для категорий строительство и хоз. принадлежности;
- самые дешёвые товары это растения и искусственные растения;
- чёткой зависимости между количеством товаров и выручкой заказа не наблюдается.

анализ сезонности:

месячная сезонность:

- наблюдается повышенное количество покупок в ноябре и декабре;
- популярность растений выше обычного с марта по май - особенно в мае. Скорее всего интерес проявляют дачники, покупая рассаду;
- заинтересованность искусственными растениями проявляется в октябре и ноябре, что может быть связано с подготовкой к новому году;
- в декабре и январе немного увеличиваются продажи кухонных, хозяйственных принадлежностей и товаров категории "Дом" что может быть связано с новогодними праздниками;
- в остальных категориях месячная сезонность не проявляется;

часовая сезонность:

- активность покупателей возрастает с 7 утра и достигает в 11 часов утра "плато", которое длится до 17 часов;
- наблюдается заметный провал активности в 18 часов, и это может быть связано с тем, что потребители находятся на пути с работы домой;
- после 18 00 клиенты проявляют меньшую активность;
- часовая сезонность в категориях не обнаружилась.

Статистические гипотезы:

Проверка сезонности доли товаров в популярной категории:

- Доля купленных товаров категории Растения с марта по май **>** доли купленных товаров с ноября по февраль и с июня по октябрь в объединении.

Проверка сезонности средней выручки товаров в популярной категории:

- Средняя выручка купленных товаров категории Растения с марта по июнь (первая группа) **<=** средней выручки купленных товаров с ноября по апрель и с июля по октябрь в объединении (вторая группа) несмотря на увеличение доли количества заказанных товаров в первой группе.
- Выделена подкатегория "Рассада", которая набирает популярность в апреле и мае и из-за дешевизны товаров уменьшает среднюю выручку всей категории "Растения".

Проверка статистической значимости различия средней выручки заказов с одним товаром со средней выручки заказов, в которых товаров больше одного:

- Средняя выручка заказов с одним товаром **<** средней выручки заказов, в которых товаров больше одного.
- Чёткой зависимости цены заказа от количества товаров нет, однако, средняя выручка с несколькими товарами в заказе всё же выше.