КОМАНДА YEET

Explorary Data Analysis

2022

ПРИЗНАКИ

Всего в датасете 149 признаков и 135061 строка данных.
Признаки условно делятся на 8 групп.

01 **clinent_id**

Идентификационный номер покупателя, является уникальным, принимает значения от 0 до 135060.

02 **rto_n***

Сумма товарооборота в рублях на покупателя в месяц, где n - номер месяца, принимающий значения от 6 до 12.

03 **rto_n_category* ****

Сумма товарооборота в рублях на покупателя в месяц n по категории category. Сумма по этим столбцам НЕ дает значение столбца rto_n.

04 **rto_std_n***

Стандартное отклонение суммы товарооборота от чека к чеку в месяц n.

05 **rto_stddev_n_category* ****

Стандартное отклонение суммы товарооборота от чека к чеку в месяц n по категории category.

06 **cnt_checks_n***

Количество чеков покупателя в месяц n.

07 **cnt_checks_n_category* ****

Количество чеков покупателя в месяц n в категории category.

08 **is_in_club**

Флаг участи я в клубе, где 1 - покупатель является членом клуба,
0 - не является. Целевая переменная.

* n - номер месяца, информация о котором представлена далее
** category - категория товара из списка категорий, представленного далее

О ДАННЫХ

Номер месяца

Номер месяца n для всех признаков принимает значения от 6 до 12.
То есть данные представлены за 7 месяцев: с июня по декабрь.

Категория товара

Всего в данных представлено 6 категорий:

- Крупы и зерновые
- Мясная гастрономия
- Овощи - Фрукты
- Птица и изделия из птицы
- Рыба и рыбные изделия
- Сыры

Каждый признак, имеющий разбиение по категориям, содержит только категории из списка выше.

Пропуски в данных

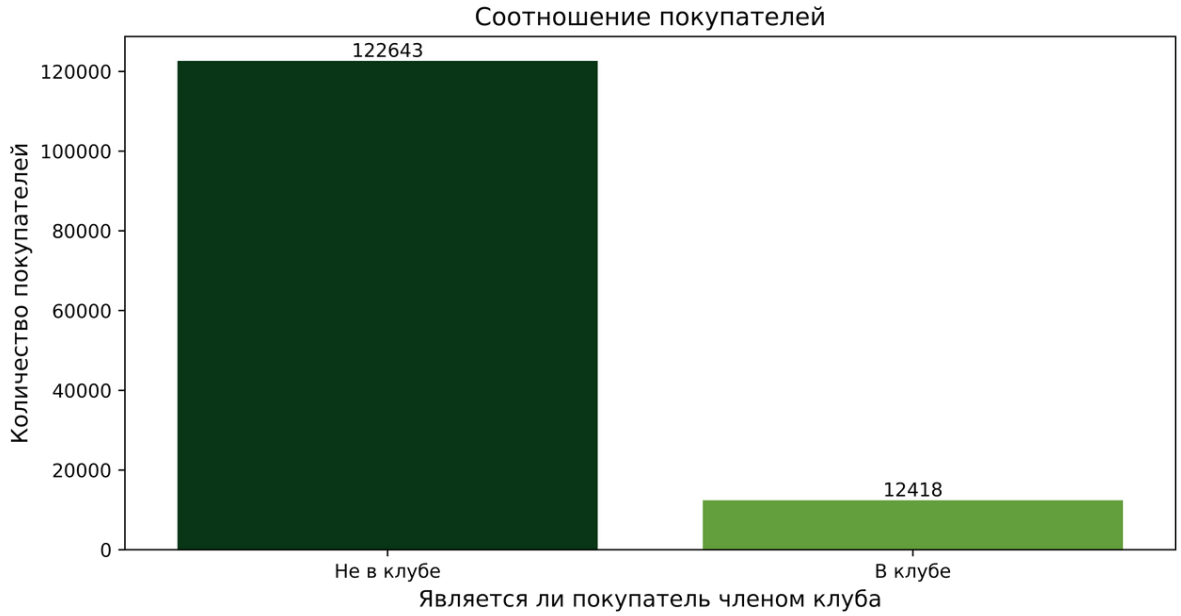
Пропущенные значения встречаются во всех столбцах, кроме client_id и is_in_club.

В столбцах групп rto_n, rto_n_category, cnt_checks_n, cnt_checks_n_category пропущенные значения интерпретируются как отсутствие покупок в данный месяц (по данной категории).

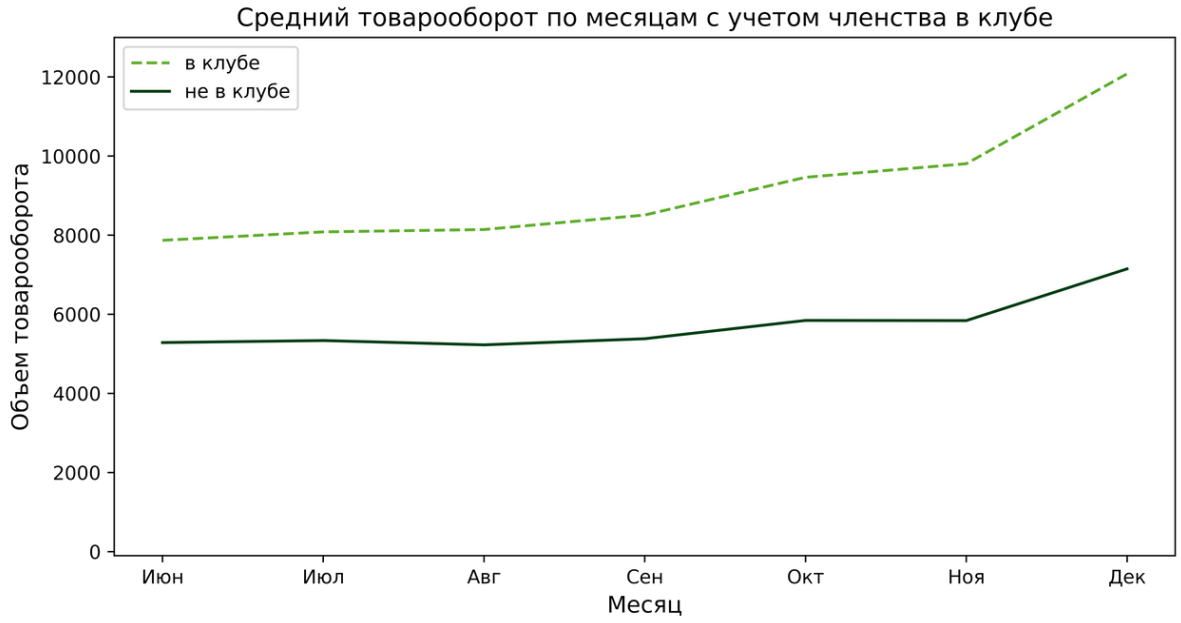
В столбцах rto_std_n и rto_stddev_n_category пропущенные значения интерпретируются двумя способами:

1. в данный месяц (по данной категории) покупки не совершались
2. в данном месяце (по данной категории) был пробит лишь один чек, откуда следует отсутствие стандартного отклонения, так как для его расчета необходимо иметь как минимум два чека.

Соотношение классов покупателей



В датасете покупателей, не вступивших в клуб, **на 82% больше**, чем уже вступивших. Члены Клуба Полезных Привычек составляют **9% от всей выборки**.

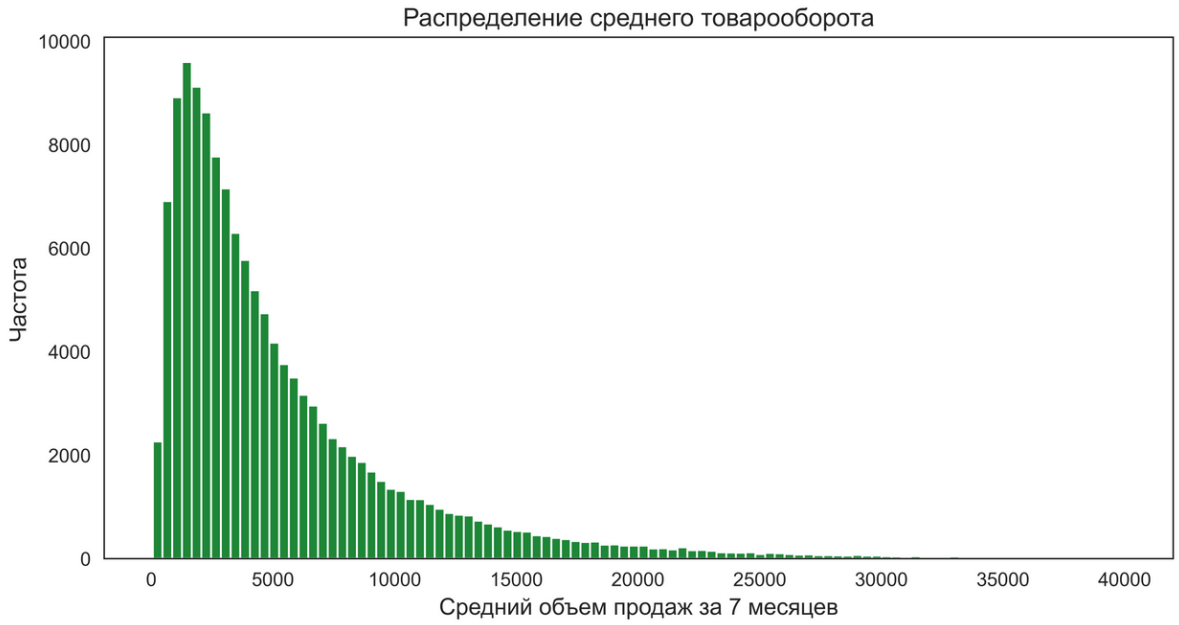


В среднем товарооборот **выше** у покупателей, состоящих в Клубе Полезных Привычек, приблизительно **на 39%**.

Минимальный товарооборот для обеих категорий приходится на июнь, **максимальный** - на декабрь.

В среднем товарооборот **увеличивается** с течением времени.

Товарооборот в месяц



На графике представлено распределение среднего товарооборота по столбцам категории rto_n.

Все столбцы категории rto_n имеют логнормальное распределение.

Значения товарооборота за один месяц:

- Минимальное: 0 руб.
- Максимальное: 794 111 руб.
- Среднее: 5 572 руб.

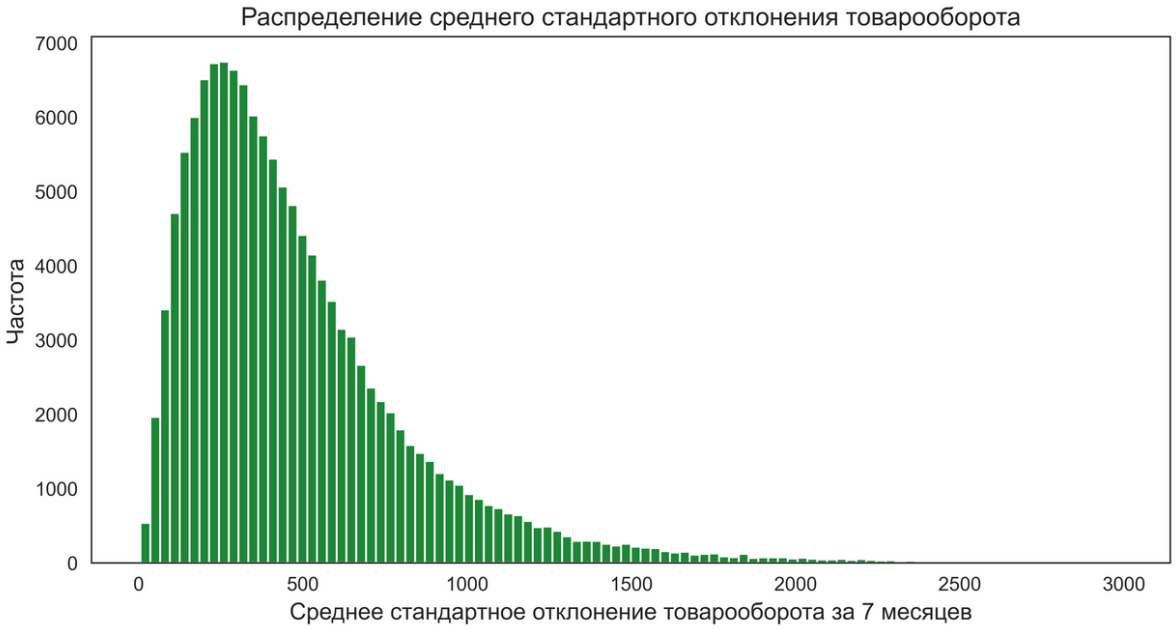
Значение среднего товарооборота в месяц за период:

- Минимальное: 62 руб.
- Максимальное: 279 976 руб.
- Медианное: 3 714 руб.

Значение суммарного товарооборота за период:

- Минимальное: 436 руб.
- Максимальное: 1 959 837 руб.
- Среднее: 39 005 руб.
- Медианное: 26 001 руб.

Стандартное отклонение товарооборота



На графике представлено распределение среднего стандартного отклонения товарооборота по столбцам категории rto_std_n.

Все столбцы категории rto_std_n также имеют логнормальное распределение.

Значения SD* товарооборота за один месяц:

- Минимальное: -2.68 руб.
- Максимальное: 24 627 руб.
- Среднее: 490 руб.

Значение среднего SD* товарооборота в месяц за период:

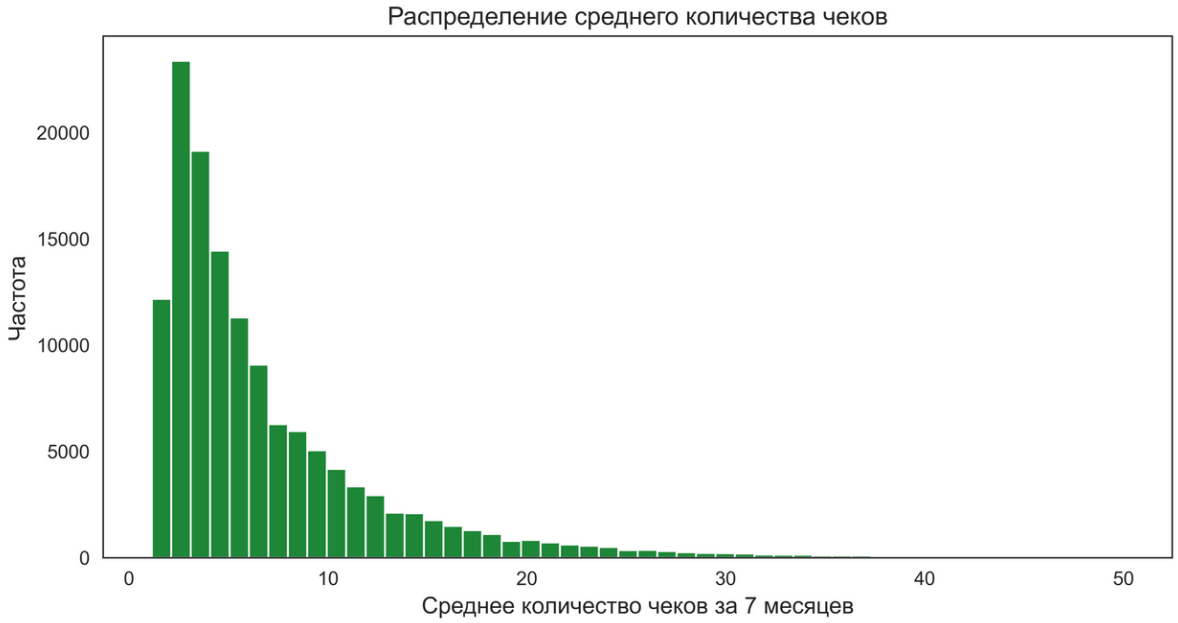
- Минимальное: 5 руб.
- Максимальное: 7 300 руб.
- Медианное: 396 руб.

Значение суммарного SD* товарооборота за период:

- Минимальное: 39 руб.
- Максимальное: 51 104 руб.
- Среднее: 3 435 руб.
- Медианное: 2 775 руб.

* - SD - standard deviation - сокращение словосочетания "стандартное отклонение"

Количество чеков в месяц



На графике представлено распределение среднего количества чеков по столбцам категории cnt_n.

Все столбцы категории cnt_n также имеют логнормальное распределение.

Количество чеков за один месяц:

- Минимальное: 0
- Максимальное: 1044
- Среднее: 7.2

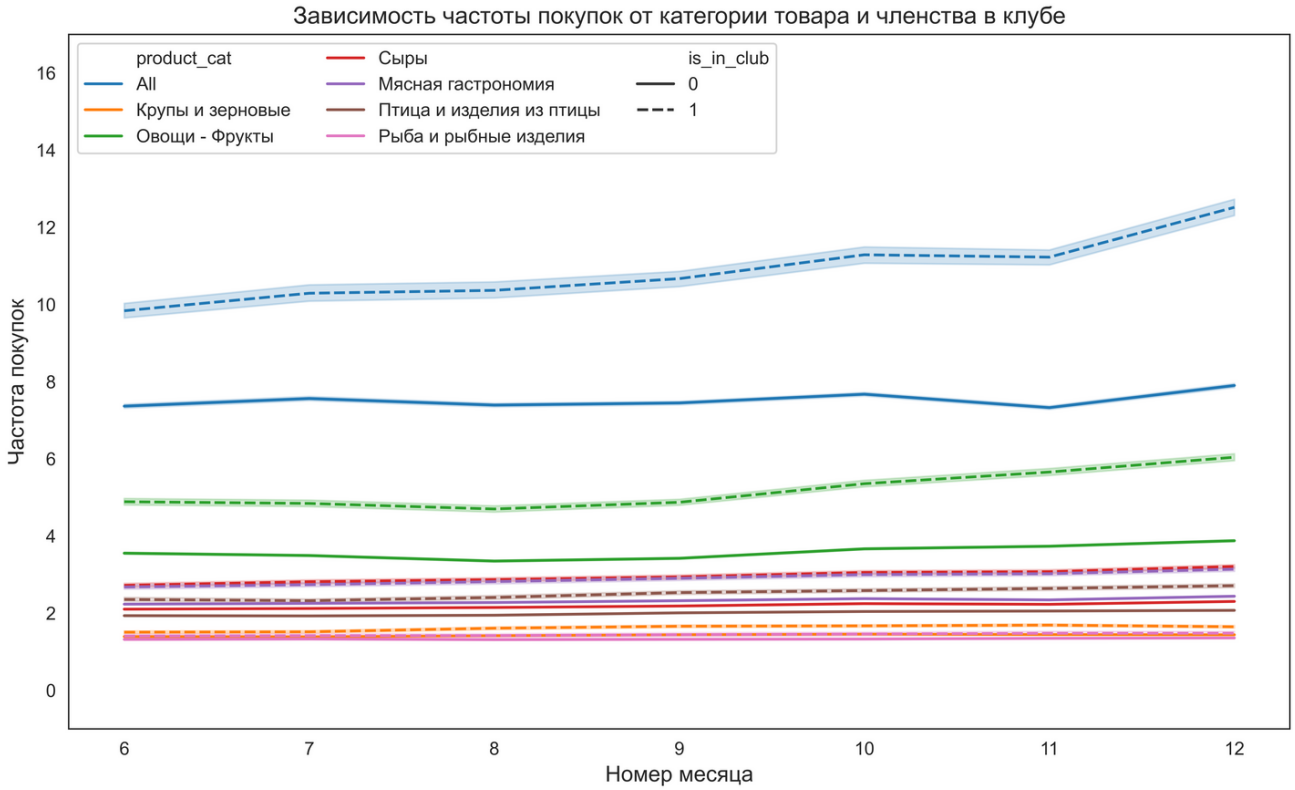
Среднее количество чеков в месяц за период:

- Минимальное: 1.2
- Максимальное: 343.5
- Медианное: 5

Суммарное количество чеков за период:

- Минимальное: 8
- Максимальное: 2 405
- Среднее: 50.5
- Медианное: 35

Категории продуктов



Самая **популярная** категория: Овощи - Фрукты.
Самая **непопулярная** категория: Рыба и рыбные изделия.

Для всех категорий справедливо, что в среднем товары **чаще покупаются членами Клуба Полезных Привычек**, чем покупателями, не вступившими в Клуб.

ВЫВОДЫ

01

Данные необходимо чистить

В данных встречаются пропущенные и отрицательные значения. Пропущенные значения заполняются нулями, а отрицательные стандартные отклонения устраняются посредством возведения в квадрат.

02

Логнормальные распределения

Все числовые признаки имеют имеют логнормальное распределение. С такими данными модель будет работать хуже, поэтому распределения приводятся к нормальным посредством применения логарифма.

03

Дисбаланс классов

Из-за того, что не вступивших в клуб сильно больше, чем тех, кто вступил, модель может показывать плохое качество, относя всех клиентов к классу не вступивших в клуб.

04

Прибыльность классов

В результате разведывательного анализа данных было выявлено, что вступившие в Клуб покупатели приносят примерно на 39% больше прибыли.