Влияние коммуникаций на клиентов банка

Нугайгулова Рушана, Томашук Анна,
Шарушкина Елизавета Э-2109

ТИНЬКОФФ

Исследовательский вопрос

Как количество коммуникаций влияет на поведение клиентов в отношении трат и предложений о кэшбэках?

Как обновление приложения повлияло на поведение клиентов банка?

1. Взаимосвязь коммуникаций и покупок с кэшбэком

Чем больше коммуникаций было, тем более вероятно, что человек совершит покупку с кэшбэком

(подтверждаем)

1. Взаимосвязь коммуникаций и покупок с кэшбэком

В столбце pur_with_cb хранятся категориальные данные:

0 - если кэшбэк нулевой, то есть совершил покупку без кэшбэка

1 - если человек получил кэшбэк, то есть совершил покупку с кэшбэком

```
df_1['pur_with_cb'] = df_1['cb_sum'].apply(lambda x: 1 if x > 0 else 0)
df_1
```

	client_id	cb_sum	cnt_communication	pur_with_cb
0	162353149	667.0	6	1
1	666605735	0.0	0	0
2	558610079	0.0	0	0
3	769974329	427.0	3	1
4	350605405	358.0	4	1
...	...	...	...	...
49995	377949465	2555.0	3	1
49996	299560244	136.0	1	1
49997	66958117	448.0	1	1
49998	549553108	488.0	3	1
49999	151873511	2747.0	2	1

1. Взаимосвязь коммуникаций и покупок с кэшбэком

Коэффициент ковариации между cnt_communication и pur_with_cb

1.070553e-01 → Связь существует, она прямая

Коэффициент корреляции Кендалла между cnt_communication и pur_with_cb

0.162729 → По шкале Чеддока связь слабая

```
st.pointbiserialr(df_1['cnt_communication'], df_1['pur_with_cb'])
```

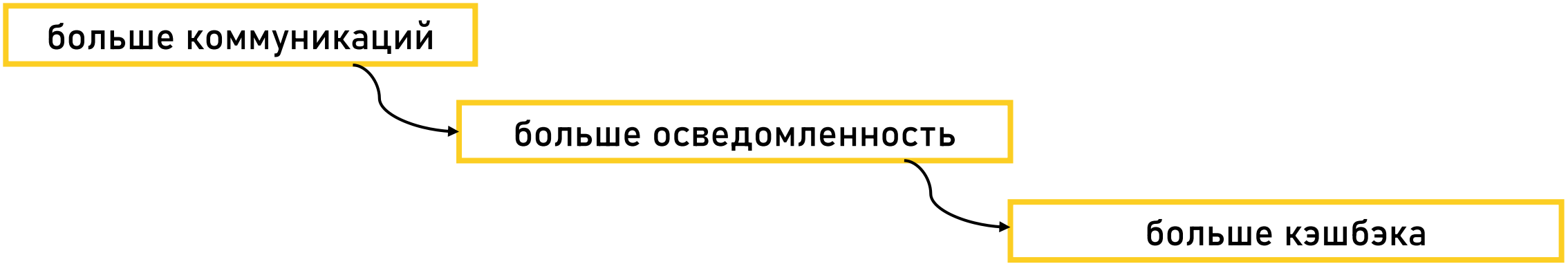
```
PointbiserialrResult(correlation=0.17312015246662177, pvalue=0.0)
```

→ p-value < 0.05, это говорит о том, что есть основания отвергнуть нулевую гипотезу с вероятностью 0.95 о том, что нет корреляции между cnt_communication и pur_with_cb.
Это означает, что есть статистически значимая корреляция между рассматриваемыми переменными.

1. Взаимосвязь коммуникаций и покупок с кэшбэком

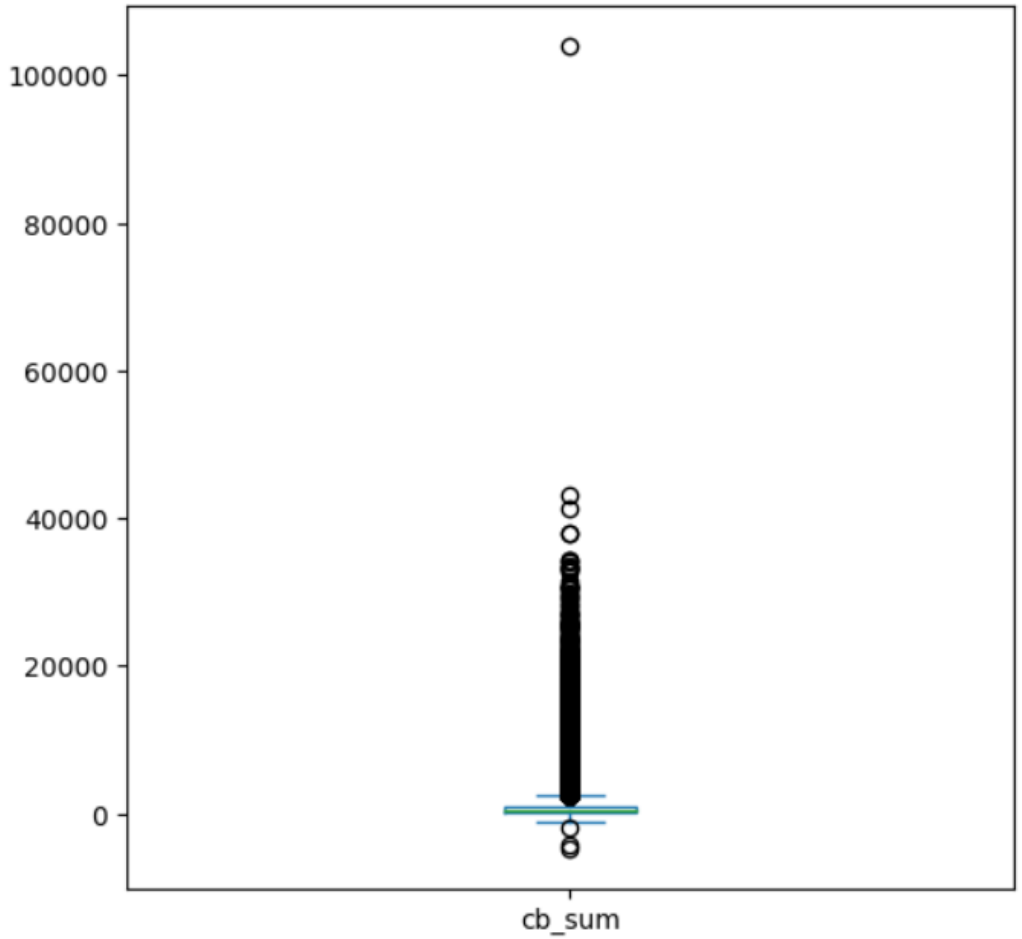
Чем больше коммуникаций было отправлено, тем больший кэшбэк получил клиент

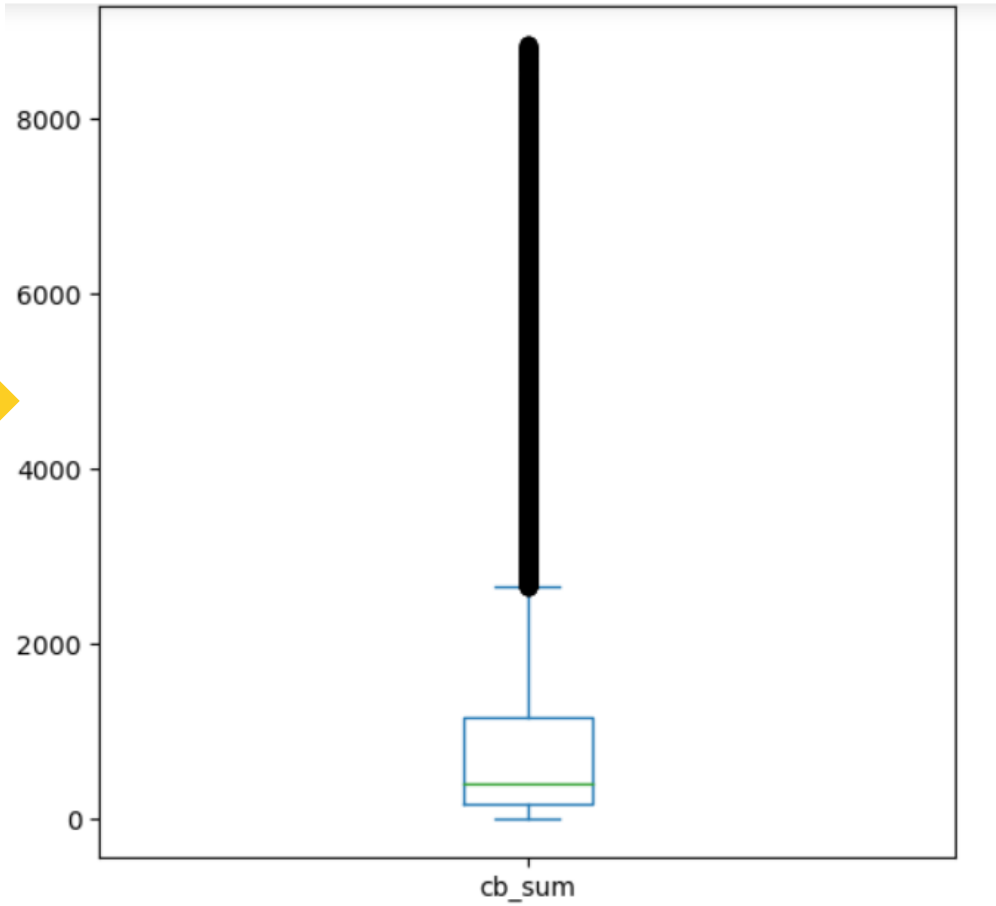
(сомнения)



1. Взаимосвязь коммуникаций и покупок с кэшбэком

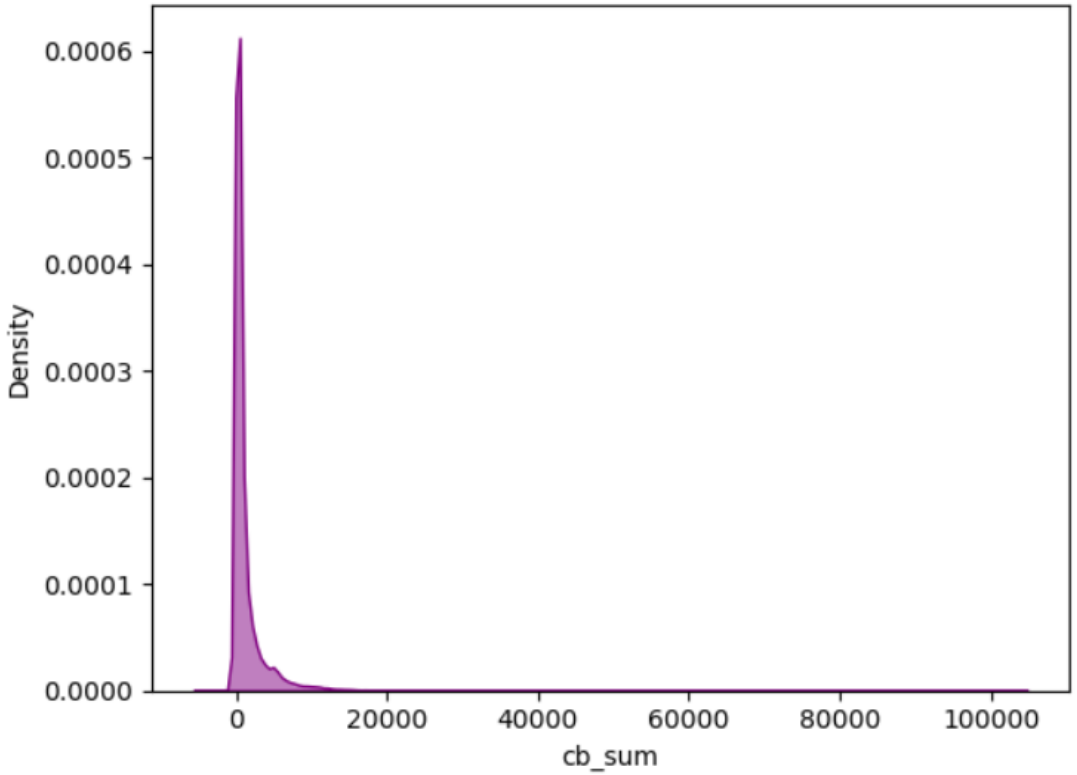
Обработка столбца cb_sum

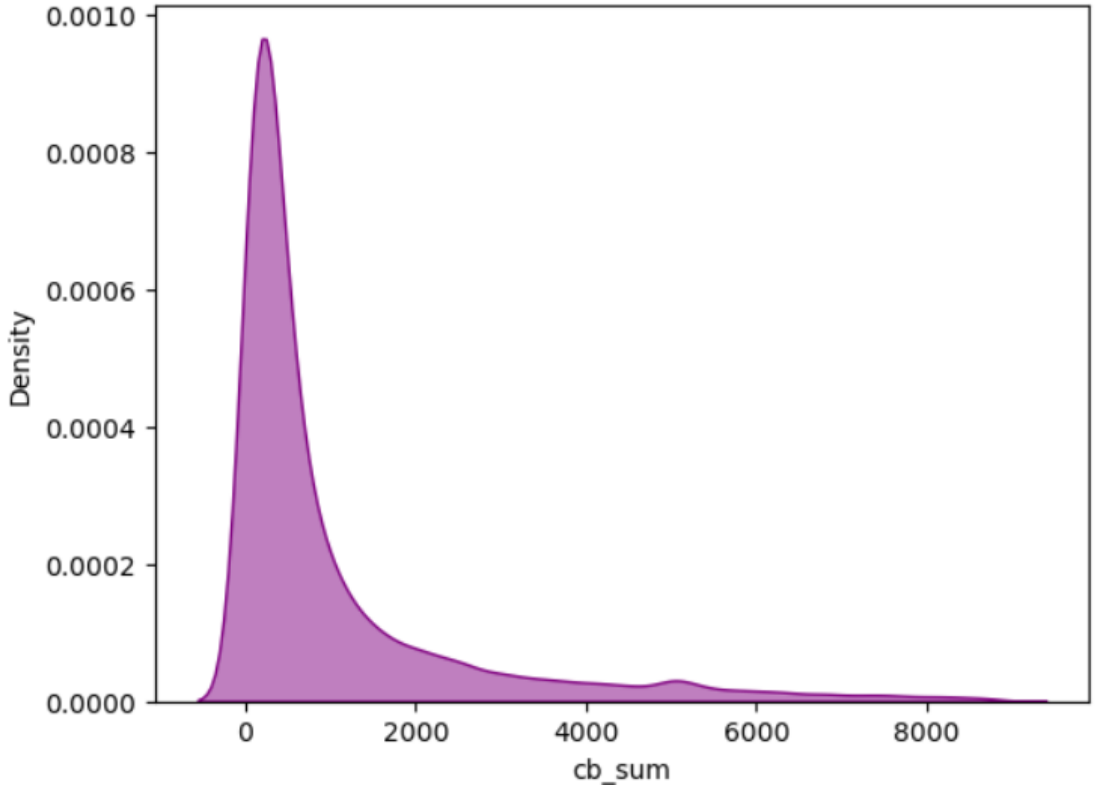




1. Взаимосвязь коммуникаций и покупок с кэшбэком

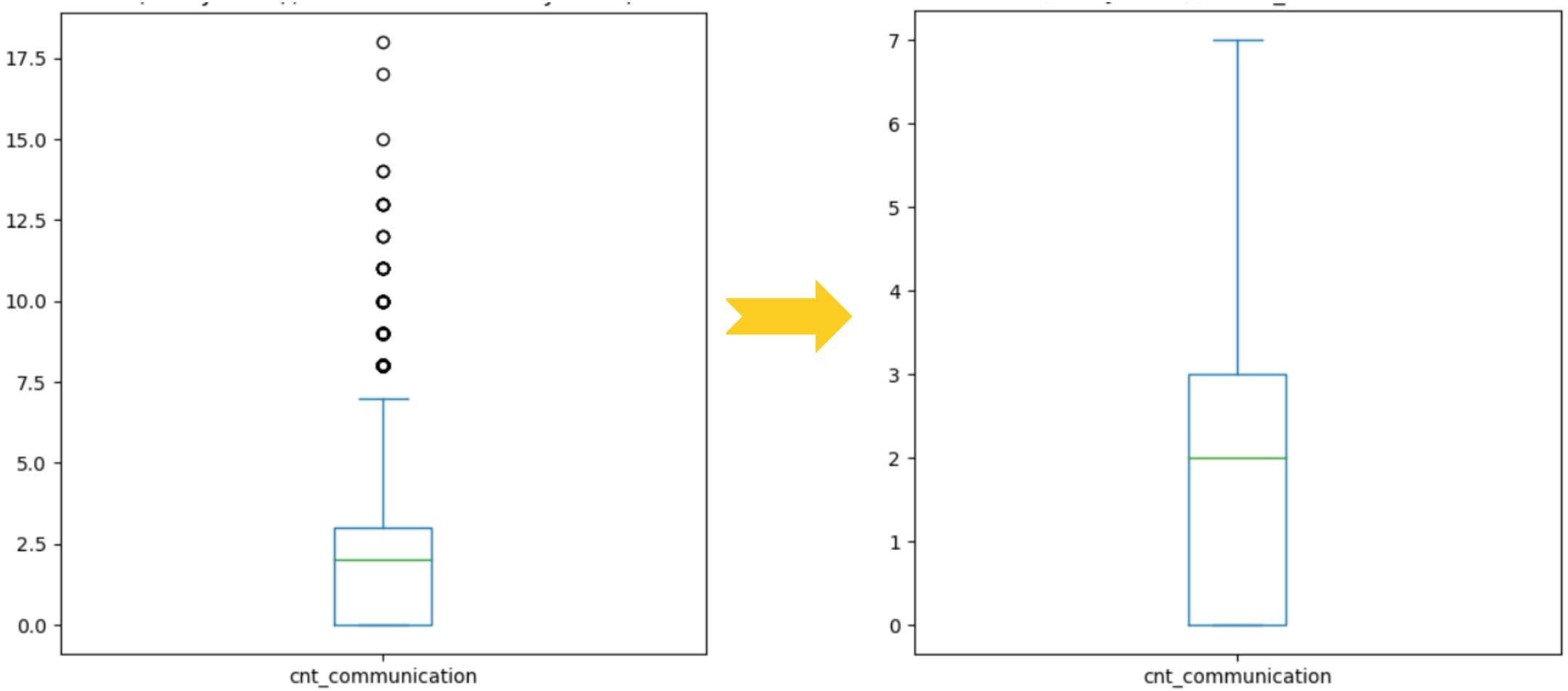
Обработка столбца cb_sum





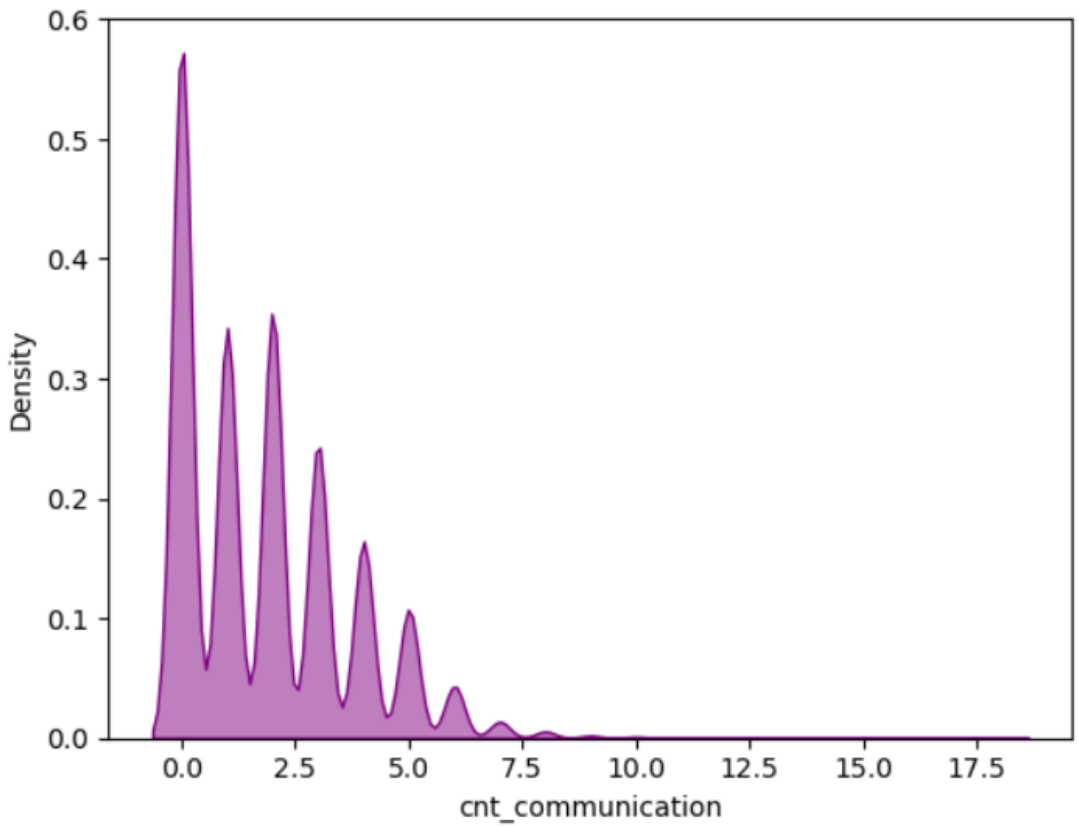
1. Взаимосвязь коммуникаций и покупок с кэшбэком

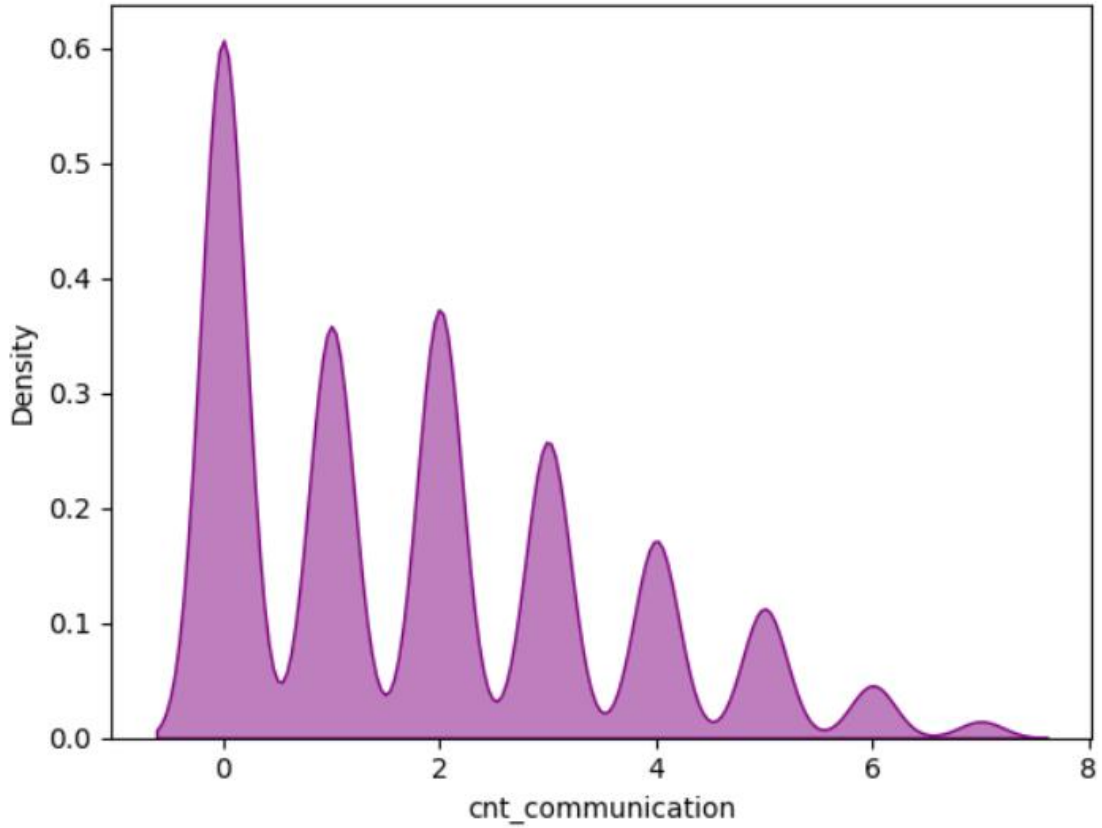
Обработка столбца cnt_communication



1. Взаимосвязь коммуникаций и покупок с кэшбэком

Обработка столбца cnt_communication





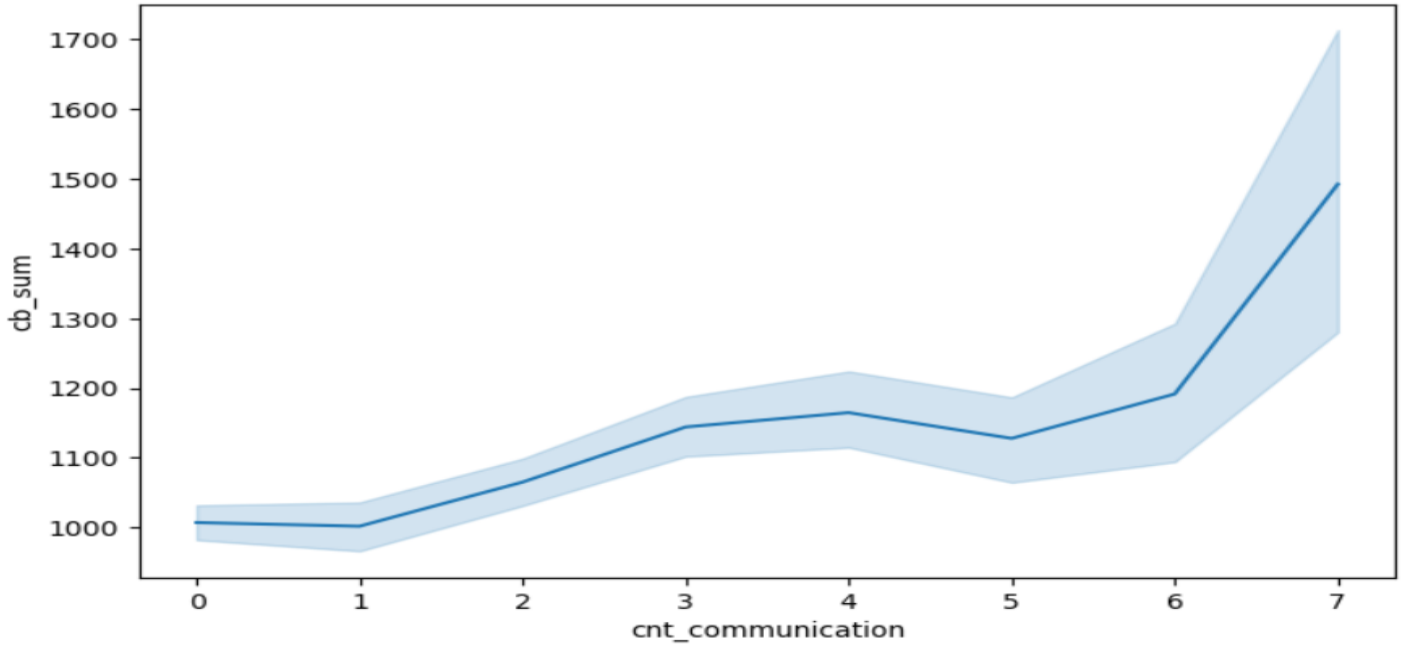
1. Взаимосвязь коммуникаций и покупок с кэшбэком

Коэффициент ковариации между cnt_communication и cb_sum

1.144702e+02 → Связь существует, она прямая

Коэффициент корреляции Кендалла между cnt_communication и cb_sum

0.051876 → По шкале Чеддока связи практически нет



2. Как взаимосвязаны обороты клиентов и выплаченный кэшбэк?

Чем больше больше сумма, потраченная клиентами, тем больше выплаченный кэшбэк

(подтверждаем)

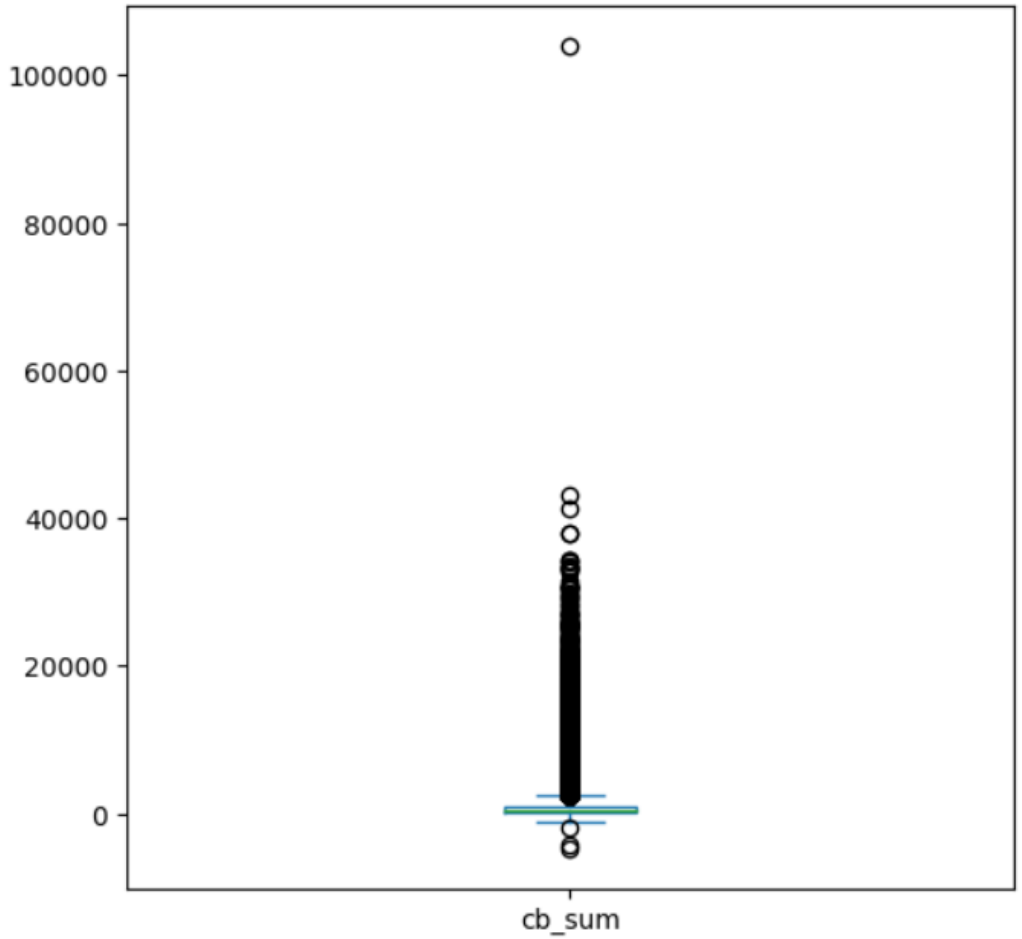
больше трат

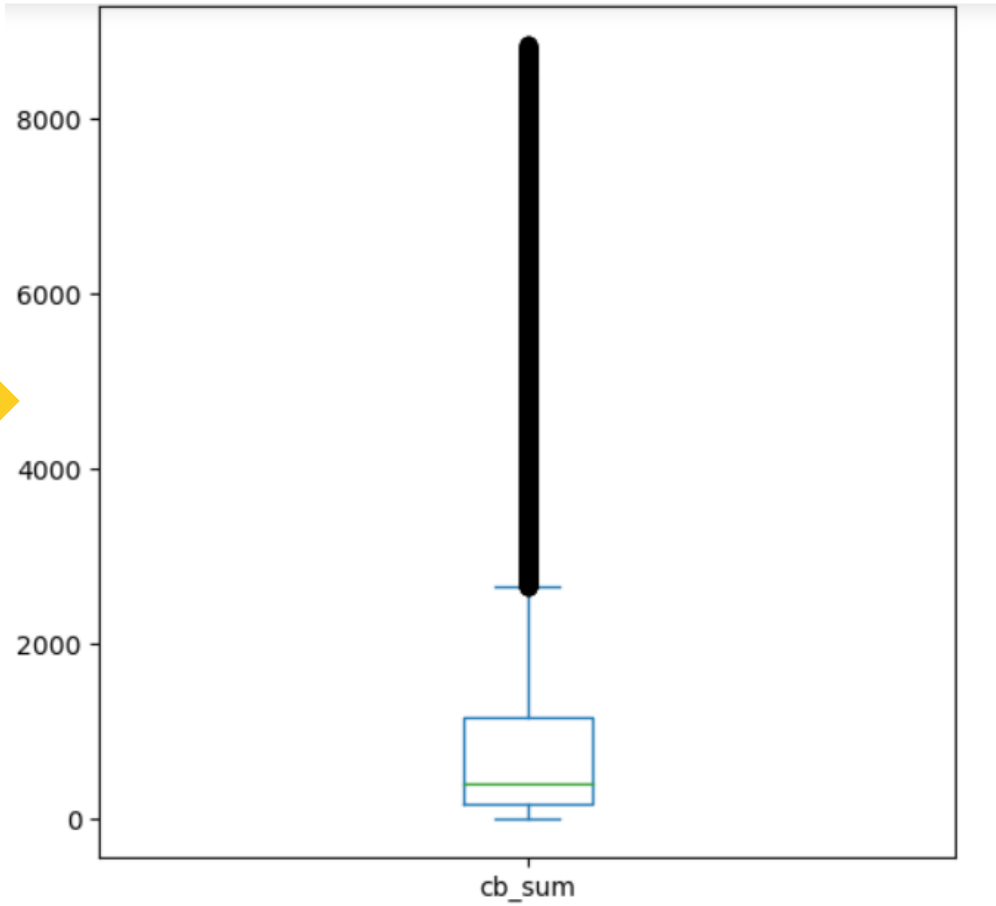
больше охват предложений кэшбэка

больше кэшбэк

2. Как взаимосвязаны обороты клиентов и выплаченный кэшбэк?

Обработка столбца cb_sum





2. Как взаимосвязаны обороты клиентов и выплаченный кэшбэк?

Коэффициент ковариации между pur_sum и cb_sum

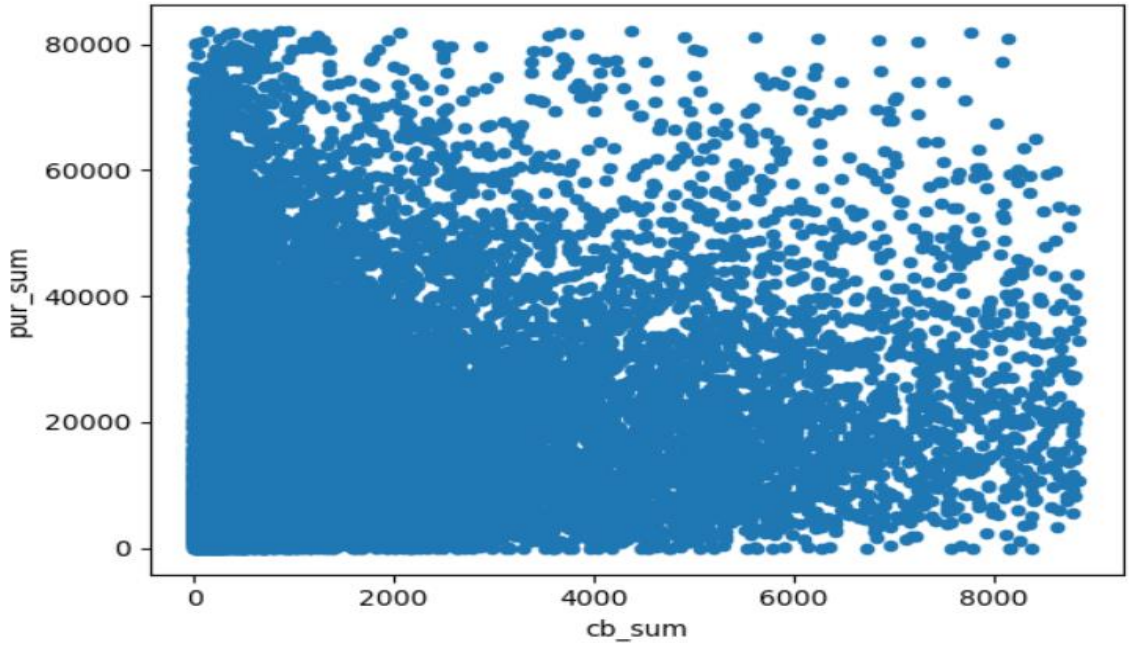
4.966327e+06

Связь существует, она прямая

Коэффициент корреляции Кендалла между pur_sum и cb_sum

0.169672

По шкале Чеддока связь слабая



3. 3 месяца назад произошло глобальное обновление приложения в т.ч. методики взаимодействия с пользователем. Необходимо оценить эффект обновления.

После обновления приложения средняя сумма покупок увеличилась

(подтверждаем)

улучшили приложение

больше заинтересованность в предложениях кэшбэка

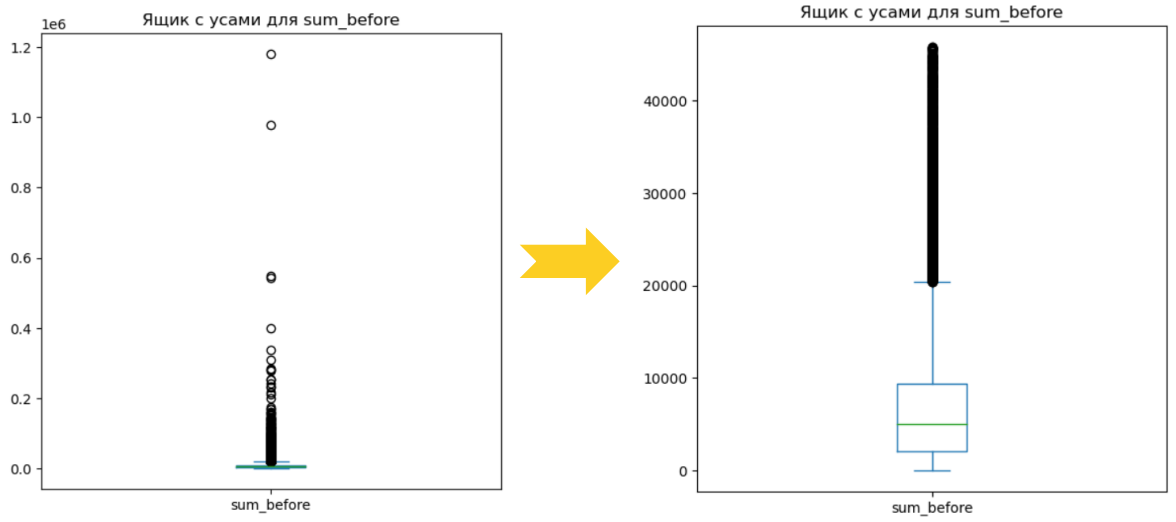
больше покупок

3. 3 месяца назад произошло глобальное обновление приложения в т.ч. методики взаимодействия с пользователем. Необходимо оценить эффект обновления.

Создали 2 датасета: до обновления (df_before) приложения и после обновления (df_last)

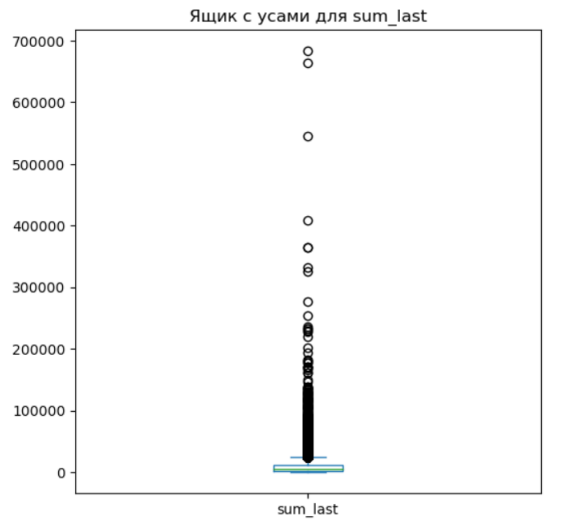
```
df_before.head()
```

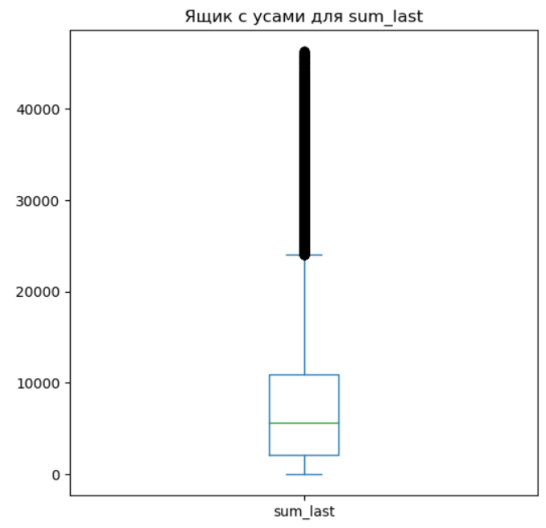
	cb_bank_before	cb_merch_before	cnt_communication	sum_before	cb_sum_before
0	125.0	119.0	6	16800.0	244.0
1	0.0	0.0	0	14.0	0.0
2	0.0	0.0	0	12474.0	0.0
3	102.0	151.0	3	904.0	253.0
4	179.0	179.0	4	4489.0	358.0




```
df_last.head()
```

	cb_bank_last_3_month	cb_merch_last_3_month	cnt_communication	sum_last	cb_sum_last
0	141.0	282.0	6	12832.0	423.0
1	0.0	0.0	0	0.0	0.0
2	0.0	0.0	0	9775.0	0.0
3	0.0	174.0	3	1164.0	174.0
4	0.0	0.0	4	7707.0	0.0





3. 3 месяца назад произошло глобальное обновление приложения в т.ч. методики взаимодействия с пользователем. Необходимо оценить эффект обновления.

Односторонний тест Уилкоксона-Манна-Уитни

H0: Средняя сумма покупок до и после обновления приложения НЕ изменилась

H1: Средняя сумма покупок до и после обновления приложения увеличилась

```
# проведем тест Уилкоксона-Манна-Уитни
alpha = 0.05
results = st.mannwhitneyu(df_last_1['sum_last'], df_before_1['sum_before'])
print('p-value', results.pvalue / 2)

if results.pvalue / 2 < alpha and df_before_1['sum_before'].mean() < df_last_1['sum_last'].mean():
    print('Отклоняем нулевую гипотезу')
else:
    print('Нет оснований отклонить нулевую гипотезу')
```

```
p-value 1.484115494963841e-34
Отклоняем нулевую гипотезу
```

Отвергаем нулевую гипотезу с вероятностью 0.95.

Средняя сумма покупок до и после обновления приложения УВЕЛИЧИЛАСЬ.

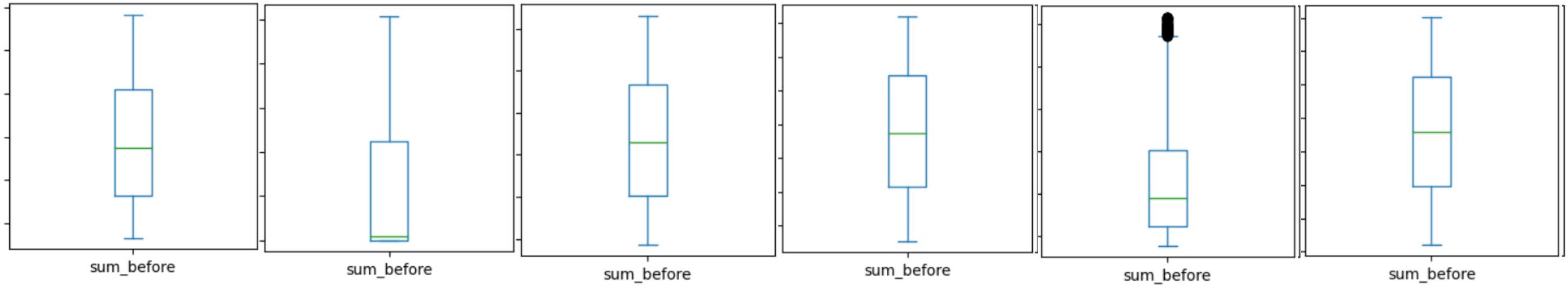
3. 3 месяца назад произошло глобальное обновление приложения в т.ч. методики взаимодействия с пользователем. Необходимо оценить эффект обновления.

Бакетный тест

Разделим датасеты до и после обновления на 15 бакетов:

```
df_before_1['bucket'] = pd.qcut(df_before_1['sum_before'], 15, labels=[i for i in range(15)])
df_before_1.head()
```

В каждом бакете данные распределены равномерно без выбросов.
Это подтверждают построенные «ящики с усами»



3. 3 месяца назад произошло глобальное обновление приложения в т.ч. методики взаимодействия с пользователем. Необходимо оценить эффект обновления.

Средняя сумма покупок до обновления приложения по 15 бакетам: 6977.02 руб.
Средняя сумма покупок после обновления приложения по 15 бакетам: 7805.41 руб.
СКО суммы покупок до обновления приложения по 15 бакетам: 798.74 руб.
СКО суммы покупок после обновления приложения по 15 бакетам: 833.61 руб.
Количество наблюдений: 15

H0: Средняя сумма покупок до и после обновления приложения НЕ изменилась

H1: Средняя сумма покупок до и после обновления приложения увеличилась

Уровень значимости: **α=0.05**

t = -7,175
df = 198

→ p_value < α →

Отвергаем нулевую гипотезу с вероятностью 0.95.

Средняя сумма покупок до и после обновления приложения ИЗМЕНИЛАСЬ.

3. 3 месяца назад произошло глобальное обновление приложения в т.ч. методики взаимодействия с пользователем. Необходимо оценить эффект обновления.

После обновления приложения сумма кэшбэка на одну отправленную коммуникацию увеличилась

(подтверждаем)

улучшили приложение

→ больше заинтересованность в предложениях кэшбэка

→ больше рублей кэшбэка приходится на 1 коммуникацию

3. 3 месяца назад произошло глобальное обновление приложения. Оценка эффекта обновления.

Для df_before и df_last рассчитали значение столбца cb_to_cnt как отношение суммы кэшбэка к отправленным коммуникациям

	cnt_communication	sum_before	cb_sum_before	cb_to_cnt
0	6	16800.0	244.0	40.666667
3	3	904.0	253.0	84.333333
4	4	4489.0	358.0	89.500000
5	2	1735.0	0.0	0.000000
6	4	7221.0	57.0	14.250000

H0: Сумма кэшбэка, приходящаяся на одну отправленную коммуникацию до и после обновления приложения НЕ изменилась

H1: Сумма кэшбэка, приходящаяся на одну отправленную коммуникацию до и после обновления приложения увеличилась

```
# проведем тест Уилкоксона-Манна-Уитни
alpha = 0.05
results = st.mannwhitneyu(df_before_2['cb_to_cnt'], df_last_2['cb_to_cnt'])
print('p-value', results.pvalue / 2)

if results.pvalue / 2 < alpha and df_before_1['sum_before'].mean() < df_last_1['sum_last'].mean():
    print('Отклоняем нулевую гипотезу')
else:
    print('Нет оснований отклонить нулевую гипотезу')
```

```
p-value 0.0
Отклоняем нулевую гипотезу
```

Отвергаем нулевую гипотезу с вероятностью 0.95.

Сумма кэшбэка, приходящаяся на одну отправленную коммуникацию до и после обновления приложения УВЕЛИЧИЛАСЬ

3. 3 месяца назад произошло глобальное обновление приложения в т.ч. методики взаимодействия с пользователем. Необходимо оценить эффект обновления.

После обновления увеличилась сумма рублей покупки к кэшбэку

(подтверждаем)

улучшили приложение

увеличилась сумма покупок и сумма выплачиваемого кэшбэка

увеличилась сумма покупки к 1 руб. кэшбэка

3. 3 месяца назад произошло глобальное обновление приложения. Оценка эффекта обновления.

Для df_before и df_last рассчитали значение столбца pur_to_cb как отношение суммы покупки к сумме кэшбэка

cb_bank_before	cb_merch_before	cnt_communication	sum_before	cb_sum_before	pur_to_cb
125.0	119.0	6	16800.0	244.0	68.852459
102.0	151.0	3	904.0	253.0	3.573123
179.0	179.0	4	4489.0	358.0	12.539106
57.0	0.0	4	7221.0	57.0	126.684211

H0: Сумма покупки, приходящаяся на 1 руб. кэшбэка до и после обновления приложения НЕ изменилась

H1: Сумма покупки, приходящаяся на 1 руб. кэшбэка до и после обновления приложения увеличилась

```
# проведем тест Уилкоксона-Манна-Уитни
alpha = 0.05
results = st.mannwhitneyu(df_before_3['pur_to_cb'], df_last_3['pur_to_cb'])
print('p-value', results.pvalue / 2)

if results.pvalue / 2 < alpha and df_before_1['sum_before'].mean() < df_last_1['sum_last'].mean():
    print('Отклоняем нулевую гипотезу')
else:
    print('Нет оснований отклонить нулевую гипотезу')
```

```
p-value 1.3822789056474435e-150
Отклоняем нулевую гипотезу
```

Отвергаем нулевую гипотезу с вероятностью 0.95.

Сумма покупки, приходящаяся на 1 руб. кэшбэка до и после обновления приложения УВЕЛИЧИЛАСЬ

4. Выдвинете собственную гипотезу на основе имеющихся данных

Клиенты, которым отправлено больше коммуникаций, совершают траты в большем количестве категорий

(подтверждаем)

больше информации отправлено

больше осведомленность

больше категорий использует

4. Выдвинете собственную гипотезу на основе имеющихся данных

Шаг 1:

Разделили клиентов на 3 группы:

1. 'не отправлено‘ , если cnt_communication =0
2. 'отправлено мало‘, если cnt_communication = 1-2
3. 'отправлено много‘, если cnt_communication > 2

```
df.loc[(df['cnt_communication'] < 1), 'group'] = 'не отправлено'
df.loc[(df['cnt_communication'] > 0) & (df['cnt_communication'] < 3), 'group'] = 'отправлено мало'
df.loc[(df['cnt_communication'] >= 3), 'group'] = 'отправлено много'
```

Шаг 2:

Посчитали количество категорий, по которым клиент совершал покупки, за два периода (num_before и num_last).

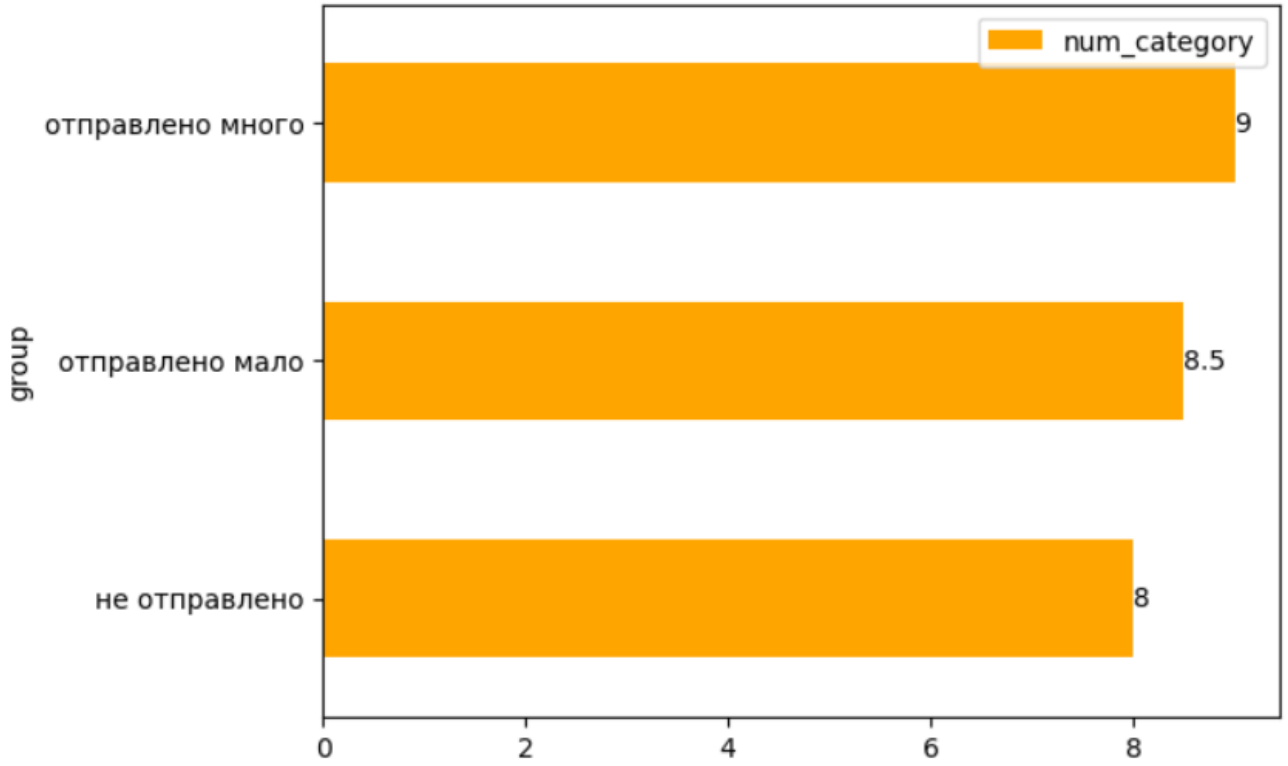
```
columns_last = df.iloc[:, 8:25].columns

for i in range(len(df)):
    last_nonzero = 0  # Переменная для подсчета ненулевых значений в каждой строке
    for column in columns_last:
        if df.loc[i, column] != 0:
            last_nonzero += 1
    df.loc[i, 'num_last'] = last_nonzero
```

Шаг 3:

Вычислили среднее значение количества используемых клиентов категорий (num_category).

group	num_last	num_before	num_category
отправлено много	11.0	13.0	12.0
не отправлено	0.0	2.0	1.0
не отправлено	9.0	9.0	9.0
отправлено много	5.0	6.0	5.5
отправлено много	8.0	10.0	9.0
...	...	...	...
отправлено много	9.0	9.0	9.0
отправлено мало	9.0	11.0	10.0
отправлено мало	8.0	8.0	8.0
отправлено много	9.0	12.0	10.5
отправлено мало	8.0	7.0	7.5

4. Выдвинете собственную гипотезу на основе имеющихся данных



По графику видно, что медианное значение количества используемых категорий для покупок растет при увеличении количества отправленных коммуникаций

Выводы по работе

После обновления приложения

- сумма кэшбэка на одну отправленную коммуникацию ↑
- средняя сумма покупок ↑
- сумма рублей покупки к кэшбэку ↑

Увеличивается количество коммуникаций

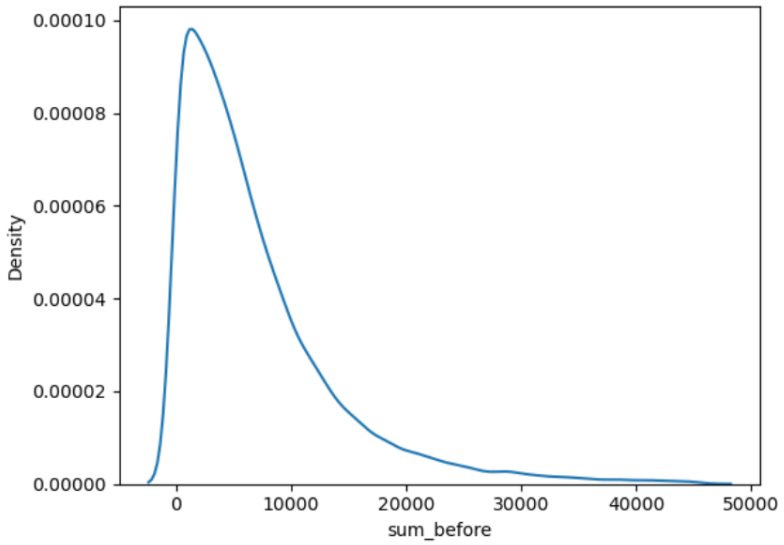
- увеличивается количество категорий покупок
- более вероятно, что человек совершит покупку с кэшбэком

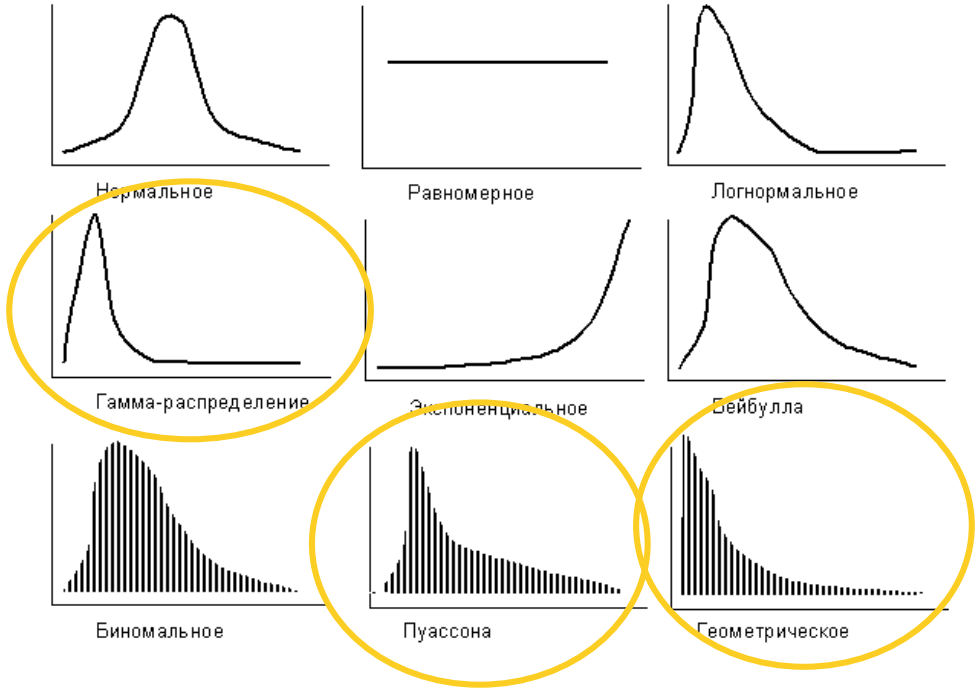
Обновление приложения хорошо сказалось на измеряемых метриках, увеличив ключевые для банка показатели

Основные проблемы в работе

- Отсутствие cnt_communication в разбивке по периодам
- Ненормальное распределение данных
- Отрицательный и нулевой кэшбек
- Скошенное вправо распределение

Наши данные





Проверяли соответствуют ли наши данные определенному виду распределения, используя тест Колмогорова-Смирнова.
Но вид распределения выявить не удалось. Во всех тестах нулевая гипотеза была отвергнута

- ## Гамма-распределение

```python
# Подгонка данных к гамма-распределению и получение параметров подгонки
shape, loc, scale = st.gamma.fit(df_before_1['sum_before'])

# Создание гамма-распределения на основе полученных параметров
gamma_dist = st.gamma(shape, loc, scale)

# Выполнение теста согласия Колмогорова-Смирнова
kstest_result = st.kstest(df_before_1['sum_before'], gamma_dist.cdf)

if kstest_result.pvalue < alpha:
    print('Отклоняем нулевую гипотезу')
else:
    print('Нет оснований отклонить нулевую гипотезу')
```

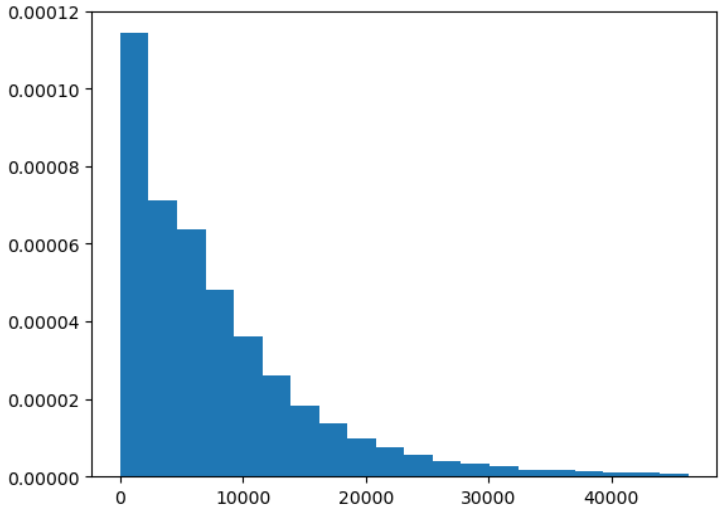
```
Отклоняем нулевую гипотезу
```

- ## Экспоненциальное распределение

```python
# Проверка гипотезы о том, что данные имеют экспоненциальное распределение
result = kstest(df_last_1['sum_last'], 'expon')
p_value = result.pvalue

if p_value < alpha:
    print('Отклоняем нулевую гипотезу')
else:
    print('Нет оснований отклонить нулевую гипотезу')
```

```
Отклоняем нулевую гипотезу
```

- ## Распределение Пуассона




```
Хи-квадрат статистика: 386565497.41537505
p-value: 0.0
Оценка параметров Пуассоновского распределения: [7859.28511799]
Результат теста Колмогорова-Смирнова: KstestResult(statistic=0.6101696603392356, pvalue=0.0, statistic_location=7585.0, statist
ic_sign=1)
```

ТИНЬКОФФ

Спасибо за внимание!