

Задача 4.

Считаем датасет. Обратим внимание, что в колонке `orders_cnt_per_user` есть значение -5. Так как число заказов не может быть отрицательным, то просто удалим эту запись о пользователе.

Хоть в таблице и есть дубликаты, но `id` пользователя нам не дан => эти дубликаты не имеют ничего общего с невалидными данными в датасете, которые не имеет смысла рассматривать

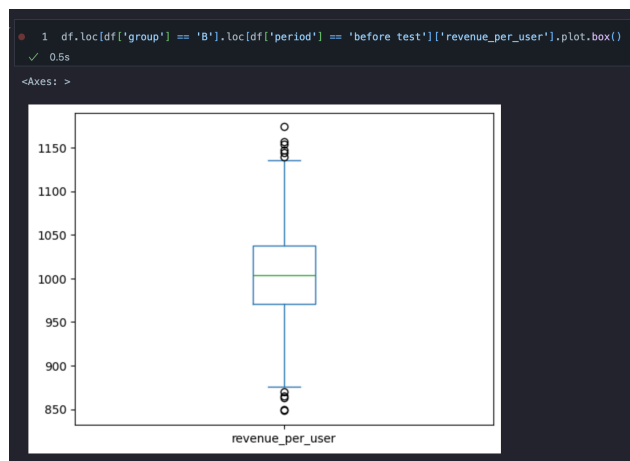
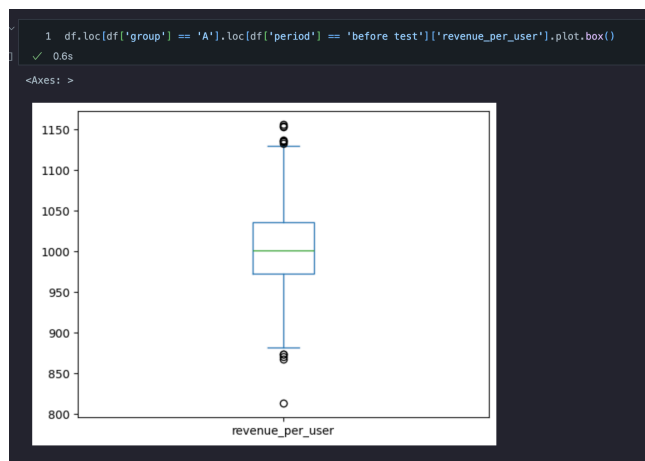
```
1 df.duplicated().value_counts()
✓ 0.0s

False    3872
True       126
Name: count, dtype: int64
```

```
1 df['orders_cnt_per_user'].value_counts()
✓ 0.0s

orders_cnt_per_user
10.0    843
9.0     741
11.0    663
8.0     528
7.0     328
12.0    319
6.0     169
13.0    127
5.0      82
14.0     63
4.0      34
15.0     25
3.0      21
16.0     19
17.0     11
2.0       9
18.0       7
1.0        5
20.0        2
19.0        2
-5.0        1
0.0         1
Name: count, dtype: int64
```

Теперь начнем исследовать данные. Для начала избавимся от выбросов по каждой из колонок (начнем с дохода), для этого построим бокс-плоты и убедимся, что выбросы действительно есть.



Избавимся от них.

```
1 agroup = df.loc[df['group'] == 'A']
2 bgroup = df.loc[df['group'] == 'B']
3
4 iqr = agroup['revenue_per_user'].quantile(0.75) - agroup['revenue_per_user'].quantile(0.25)
5 agroup = agroup.loc[(agroup['revenue_per_user'] >= agroup['revenue_per_user'].quantile(0.25) - 1.5 * iqr) & (agroup['revenue_per_user'] <= agroup['revenue_per_user'].quantile(0.75) + 1.5 * iqr)]
6
7 iqr1 = bgroup['revenue_per_user'].quantile(0.75) - bgroup['revenue_per_user'].quantile(0.25)
8 bgroup = bgroup.loc[(bgroup['revenue_per_user'] >= bgroup['revenue_per_user'].quantile(0.25) - 1.5 * iqr1) & (bgroup['revenue_per_user'] <= bgroup['revenue_per_user'].quantile(0.75) + 1.5 * iqr1)]
9
✓ 0.0s
```

Аналогично избавимся от выбросов по количеству покупок.

Теперь начнем смотреть описательные характеристики в обеих группах ДО и ВО ВРЕМЯ теста.

```
1 df.describe()
```

✓ 0.0s

	revenue_per_user	orders_cnt_per_user	converted_from_main_screen_to_item_card_screen
count	1909.000000	1909.000000	1909.000000
mean	1003.298795	9.047145	0.125720
std	46.777356	2.224996	0.331621
min	877.500000	4.000000	0.000000
25%	971.900000	8.000000	0.000000
50%	1002.200000	9.000000	0.000000
75%	1035.600000	11.000000	0.000000
max	1129.700000	15.000000	1.000000

Данные до теста
после избавления от
выбросов по столбцам
revenue_per_user и
orders_cnt_per_user

```
1 agroup.loc[agroup['period'] == 'in test'].describe()
```

✓ 0.0s

	revenue_per_user	orders_cnt_per_user	converted_from_main_screen_to_item_card_screen
count	993.000000	993.000000	993.000000
mean	1003.677845	9.884189	0.138973
std	48.700673	1.999416	0.346092
min	870.200000	4.000000	0.000000
25%	968.200000	9.000000	0.000000
50%	1006.300000	10.000000	0.000000
75%	1038.600000	11.000000	0.000000
max	1137.800000	15.000000	1.000000


```
1 bgroup.loc[bgroup['period'] == 'in test'].describe()
```

✓ 0.0s

	revenue_per_user	orders_cnt_per_user	converted_from_main_screen_to_item_card_screen
count	979.000000	979.000000	979.000000
mean	995.407252	9.955056	0.171604
std	49.517028	1.092839	0.377228
min	864.900000	6.000000	0.000000
25%	961.500000	9.000000	0.000000
50%	996.600000	10.000000	0.000000
75%	1027.850000	11.000000	0.000000
max	1131.300000	13.000000	1.000000

Теперь сравним
данные по каждой
группе в сравнении с
периодом до теста и
между собой во время
теста.

В А группе
описательные
характеристики по
“доходу с
пользователя”
практически не
отличаются (в
пределах
погрешности) от
данных до теста.

В группе Б же остались неизменными среднее значение по заказам, средний доход с пользователя немного уменьшился, увеличилась доля игроков, перешедших на экран с товарами (почти на 5%) , но это увеличение не принесло никаких изменений по другим метрикам.

Из этого можно сделать вывод, что видимого результата и изменения в поведении пользователей внедрение нового магазина не принесло (а скорее в каких-то значениях даже в отрицательную сторону отклонились значения), а так как на внедрение вероятнее всего было потрачено немало времени и денег, то лучше не стоит продолжать внедрение, а сконцентрироваться на других аспектах игры, которые смогут привлечь больше пользы компании на дистанции.