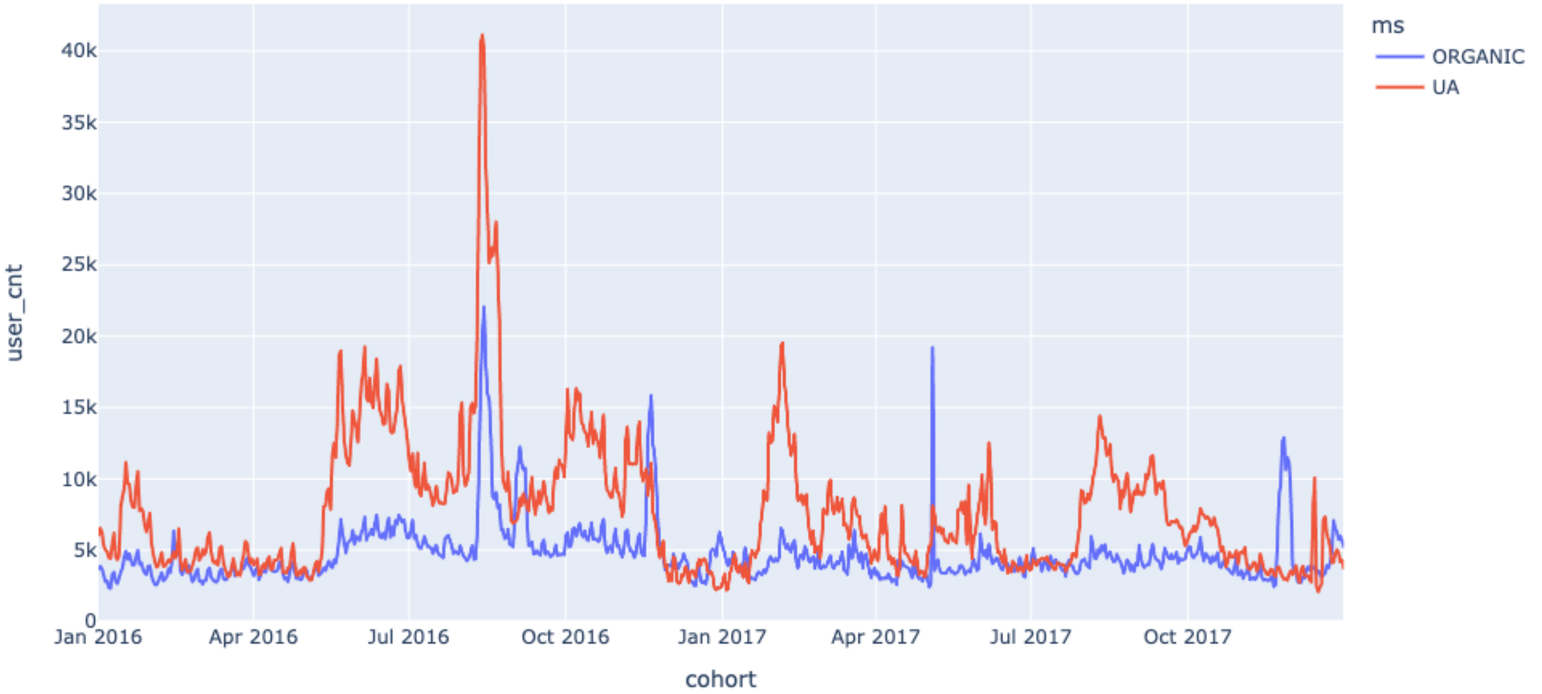
Тестовое задание

Анализ виральности игры, на основании данных инсталлов и гросса когорт

1.Вычисление “k-фактора” для трафовых пользователей

Построим график cohort от user_cnt:

Зависимость user_cnt от cohort



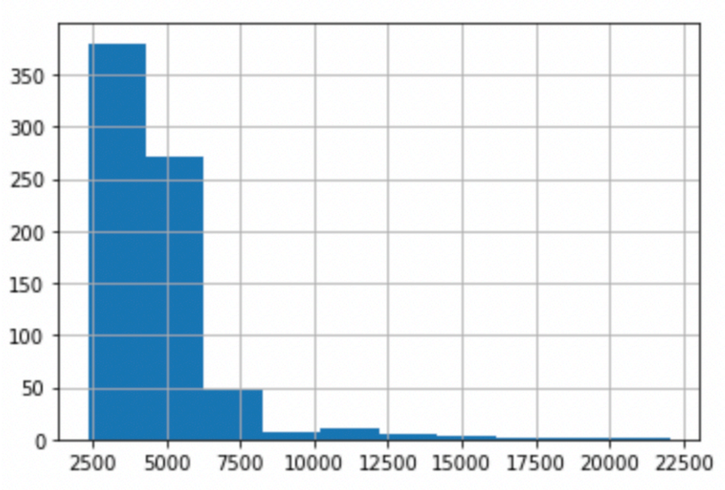
Здесь можно заметить некоторую взаимосвязь между графиками. Вместе с ростом числа трафовых растет и число органических пользователей. Сдвиг когорт ORGANIC относительно UA не наблюдается.

А значит должна существовать некоторая корреляции между временными рядами.

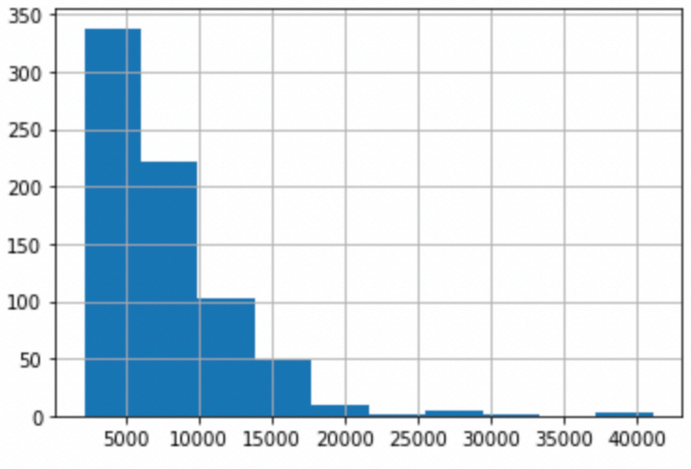
1.Вычисление “k-фактора” для трафовых пользователей

Построим теперь гистограммы распределения user_cnt UA и ORGANIC, чтобы выяснить их вид распределения:

распределение ORGANIC



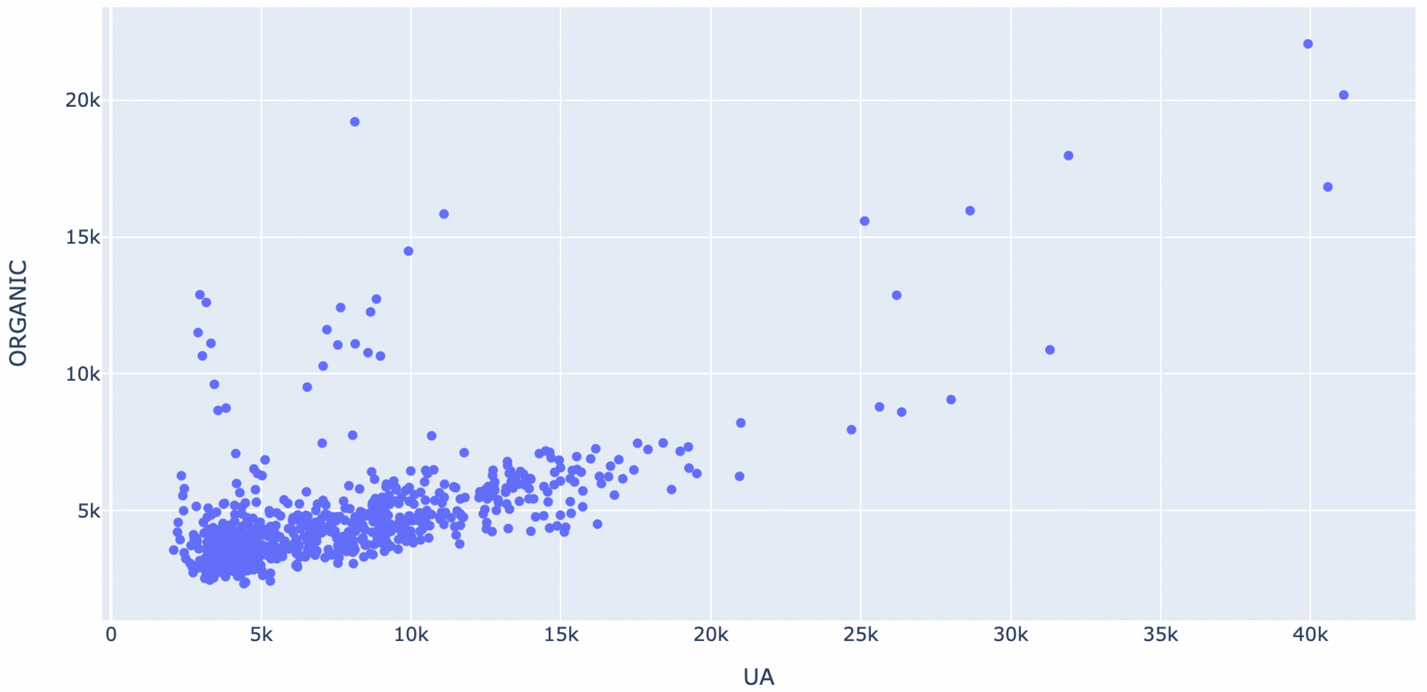
распределение UA



И для UA, и для ORGANIC - user_cnt ненормально распределены. Поэтому для оценки корреляции между ними не стоит применять коэффициент Пирсона. Используем для этой цели ранговый коэффициент Спирмена. Но перед оценкой корреляции сначала посмотрим на график зависимости UA от ORGANIC на предмет очевидных трендов и статистических выбросов

1.Вычисление “k-фактора” для трафовых пользователей

Зависимость user_cnt UA от user_cnt ORGANIC



На графике заметен линейный тренд. А еще заметно выделяются статистические выбросы ORGANIC, которые стоит рассмотреть отдельно, но для будущего регрессионного анализа их стоит отфильтровать. Поэтому уберем верхний 5% процентиль и затем оценим корелляцию с помощью коэффициента Спирмена.

1.Вычисление “k-фактора” для трафовых пользователей

Подсчет коэффициента корреляции Спирмена для usr_cnt

ms	ORGANIC	UA
ms		
ORGANIC	1.000000	0.673576
UA	0.673576	1.000000

Значение коэффициента корреляции в 0.6735 указывает на достаточно сильную корреляцию. Более того, на графике зависимости user_cnt ORGANIC от user_cnt UA наблюдался линейный тренд. Поэтому для вычисления зависимости между ними используем линейную регрессию

После проведения регрессивного анализа получился следующий результат:

```
a = 0.2011, α=5% [0.1864 - 0.2158]
b = 2849.4549, α=5% [2728.5363 - 2970.3736]
```

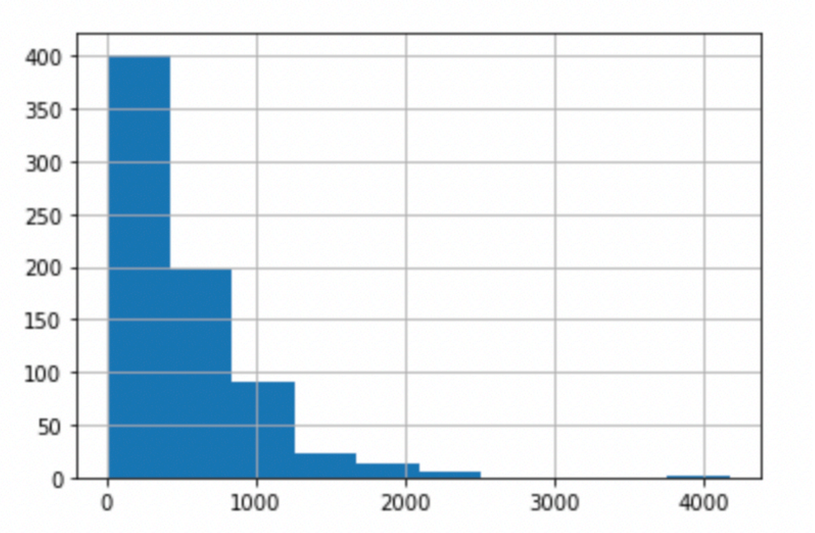
Здесь a - коэффициент прямой линейной зависимости между числом трафовых пользователей и числом органических пользователей, b - независимое от маркетинга число органических игроков.

По результатам анализа **каждый трафовый пользователь приводит в среднем 0.2 "виральных" пользователя**. Доверительный интервал полученного итогового значения k-factor'a - [0.186 - 0.216]

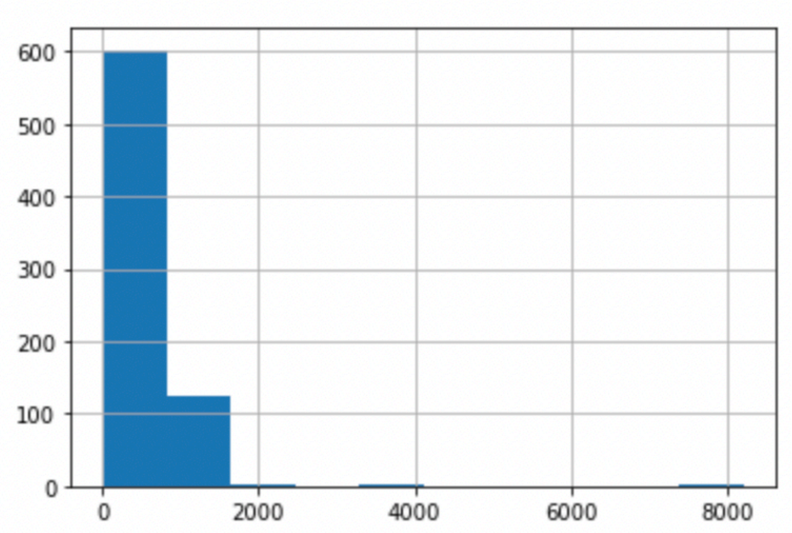
2. Вычислить k-factor для денег

Построим теперь графики распределения гроссов

Распределение UA gross



Распределение ORANIC gross

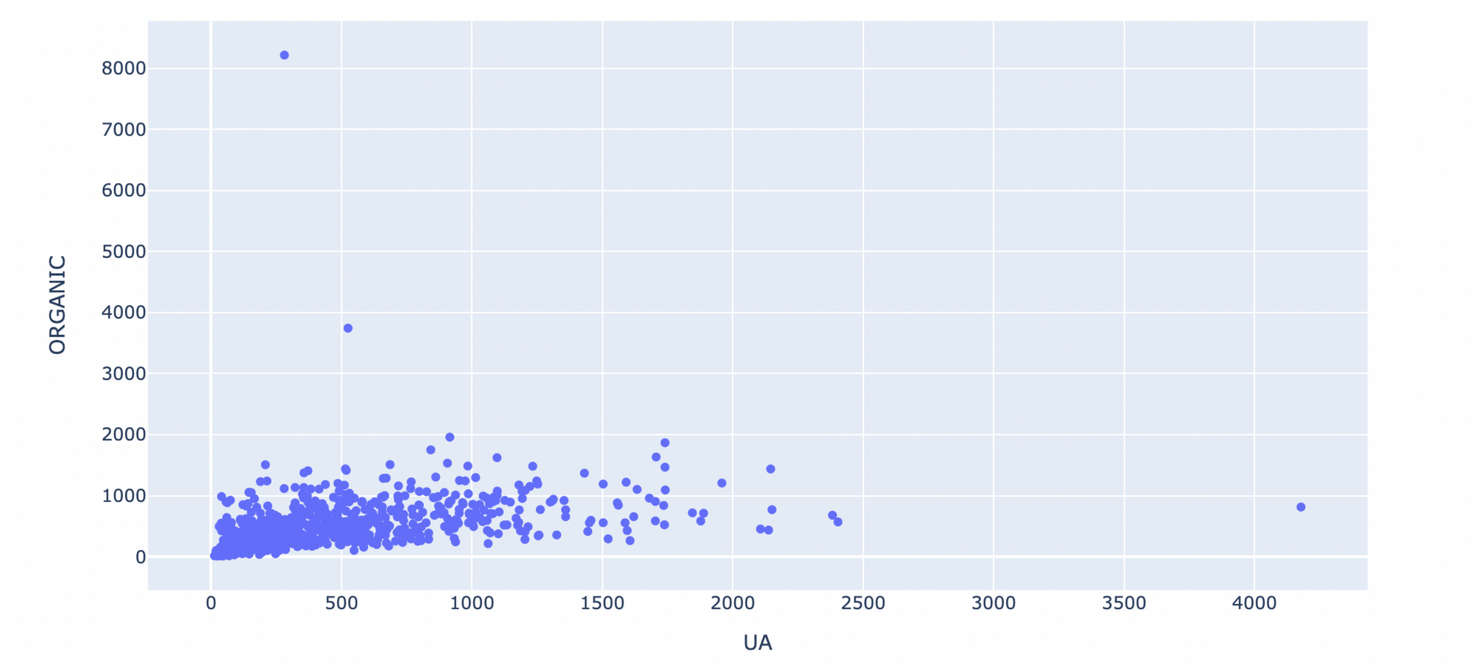


Снова видим, что гистограммы не похожи на нормальное распределение.
Значит корреляцию придется измерять с помощью коэффициента Спирмена.

2. Вычислить k-factor для денег

Построим теперь график зависимости гросса бесплатных пользователей от гросса платных

График зависимости ORANIC gross от UA gross



На графике снова наблюдаются статистические выбросы, но они незначительные. Более того, снова заметен линейный тренд.

2. Вычислить k-factor для денег

Подсчет коэффициента корреляции Спирмена для gross

ms	ORGANIC	UA
ms		
ORGANIC	1.000000	0.584476
UA	0.584476	1.000000

Корелляция не такая сильная, как в случае с пользователями. Тем не менее, функции непрерывные, тренд наблюдался линейный, корреляция достаточно значительная, поэтому попробуем применить регрессионный анализ и в этом случае

После проведения регрессивного анализа получился следующий результат:

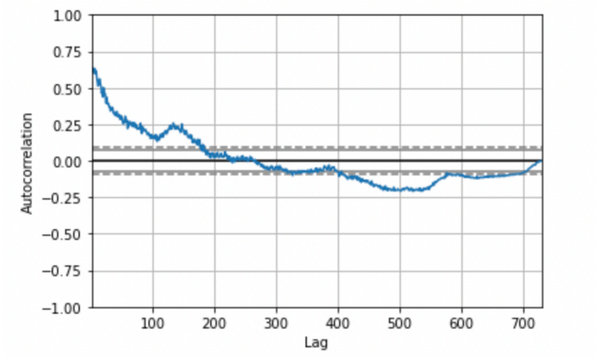
Результат: каждый доллар трафового пользователя привлекает примерно 0.34 доллара из-за виральности. Доверительный интервал [0.298 - 0.391]

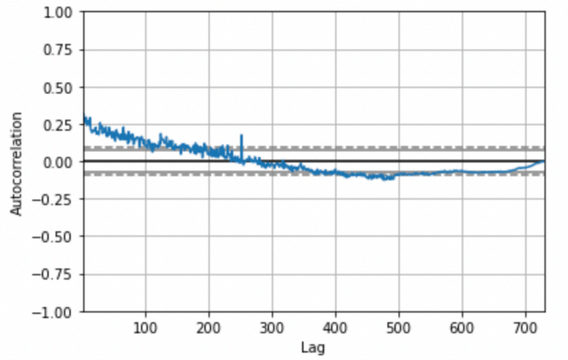
```
a = 0.3444, α=5% [0.2980 - 0.3908]
b = 331.7524, α=5% [300.9139 - 362.5910]
```

3. Дополнительный анализ данных

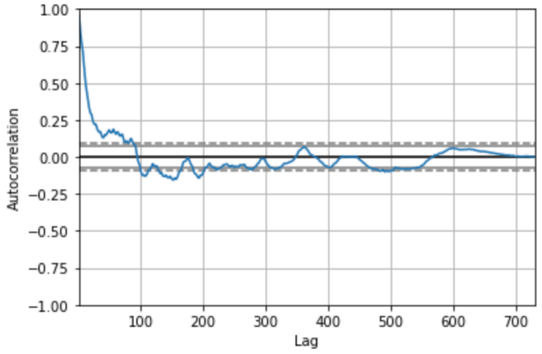
Анализ автокорреляции данных

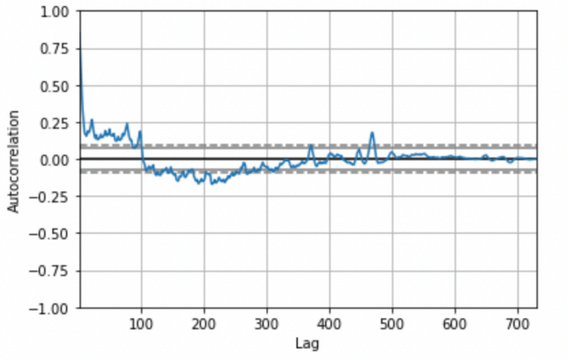
Автокорелляция UA и ORGANIC gross соответственно





Автокорелляция UA и ORGANIC user_cnt соответственно





На графиках автокорреляции заметно, что для UA свойственна значительная автокорреляция с лагом в 1 день. Это может говорить о том, что маркетинг предыдущего дня значительно влияет на результаты в следующем дне.

Кроме того, для UA gross есть небольшой рост автокорреляции с лагом на 120-130 день. Возможно у гросса есть некоторая квартальная сезонность, игроки имеют небольшую тенденцию вести себя одинаково каждый квартал

3. Дополнительный анализ данных

Анализ синхронности данных

```
synch = 0
for i in range(0, 50):
    if i == 0:
        max_corr, p = spearmanr(list(df.loc[df.ms == 'UA'].gross), list(df.loc[df.ms != 'UA'].gross))
    else:
        corr, p = spearmanr(list(df.loc[df.ms == 'UA'].gross[:-i]), list(df.loc[df.ms != 'UA'].gross[i:]))
        if corr > max_corr:
            max_corr = corr
            synch = i
```

```
synch
```

```
0
```

```
synch = 0
for i in range(0, 50):
    if i == 0:
        max_corr, p = spearmanr(list(df.loc[df.ms == 'UA'].user_cnt), list(df.loc[df.ms != 'UA'].user_cnt))
    else:
        corr, p = spearmanr(list(df.loc[df.ms == 'UA'].user_cnt[:-i]), list(df.loc[df.ms != 'UA'].user_cnt[i:]))
        if corr > max_corr:
            max_corr = corr
            synch = i
```

```
synch
```

```
0
```

Возникло предположение, что максимальная корреляция между UA и ORGANIC может возникнуть с некоторым сдвигом относительно оси cohorts. Т.е. сарафанное радио срабатывает с некоторым лагом, бесплатные пользователи приходят / приносят гросс чуть позже, чем платные. Был проведен анализ значений корреляционного коэффициента Спирмена, сдвигая данные ORGANIC от 0 до 50 дней относительно значений UA для gross и user_cnt.

В результате выяснилось, что наибольшая корреляция между данными возникает **при отсутствии сдвига когорт**.