

# Сегментация пользователей приложения «Ненужные вещи»

Выделение групп пользователей на основе их поведения на  
октябрь 2019

October 8, 2022

Руслан Мухаметшин  
[t.me/rusmux](https://t.me/rusmux)

## Цель

Выделить группы пользователей на основе их поведения и посмотреть, как группы отличаются по:

- Удержанию пользователей
- Конверсии в целевое действие (просмотр контактов)
- Времени, проводимом в приложении

## Краткие итоги

- Приложение может терять популярность.
- Кластер с 16% пользователей имеет самое высокое удержание и конверсию в просмотр контактов.
- Между всеми кластерами есть статистически значимая разница в конверсии в просмотр контактов.
- Конверсии в просмотр контактов различаются между источником Google/Яндекс и остальными источниками. Между самими источниками Google и Яндекс нет статистически значимой разницы в конверсии.

# План

- **Описание данных**

Распределение клиентов и событий

- **Исследовательский анализ данных**

Удержание пользователей, сессии, профили пользователей

- **Сегментация пользователей**

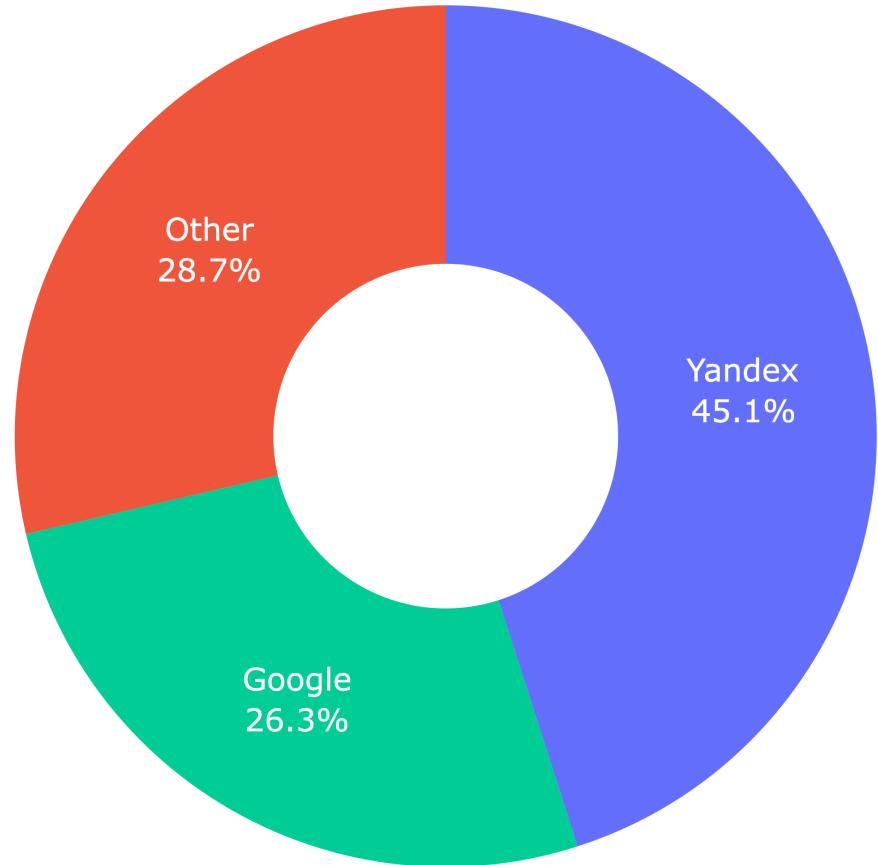
Агломеративная кластеризация, метод k-средних, PCA, UMAP

- **Проверка гипотез**

Конверсия между источниками, конверсия между кластерами

## Клиенты

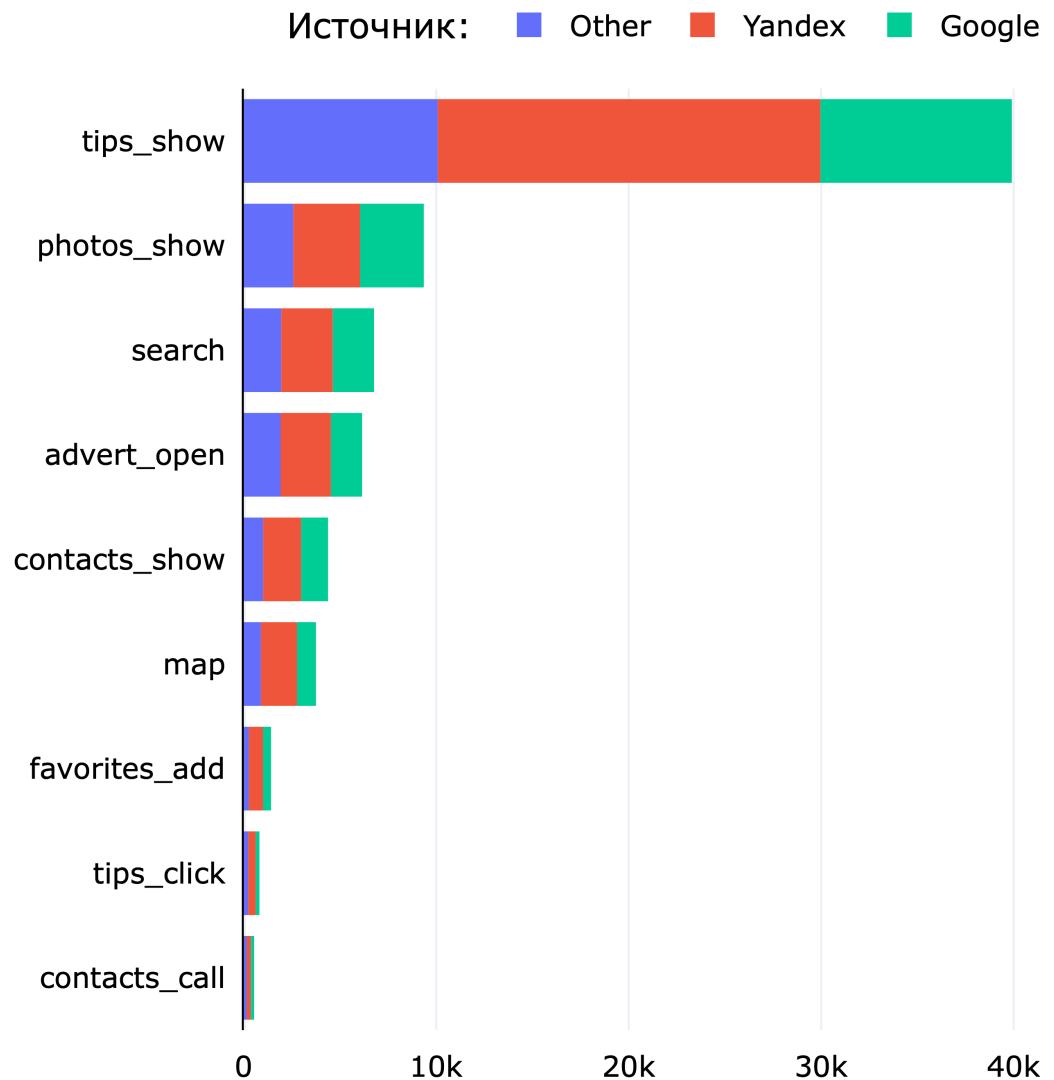
Почти половина пользователей пришли с Яндекса, а остальные примерно равномерно с Google и с других источников.



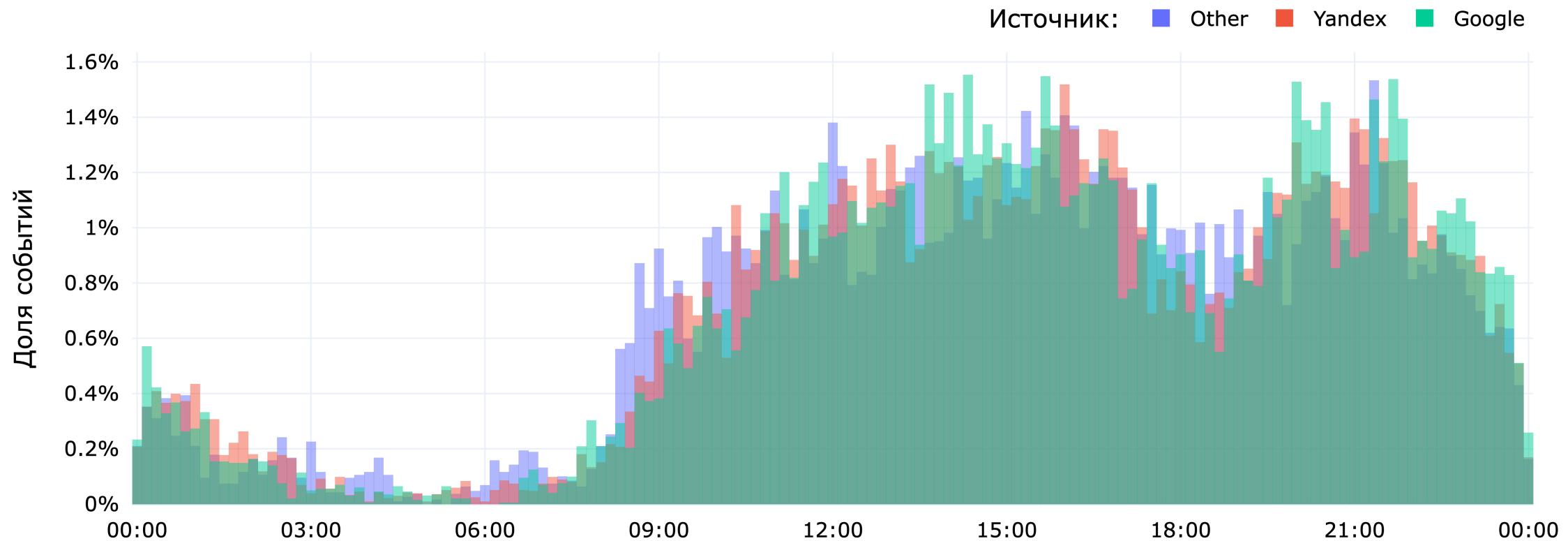
## События

Больше половины событий приходится на просмотр рекомендованных объявлений.

После них идут просмотр фотографий в карточке объявления, поиск объявлений, открытие самого объявления и просмотр контактов.



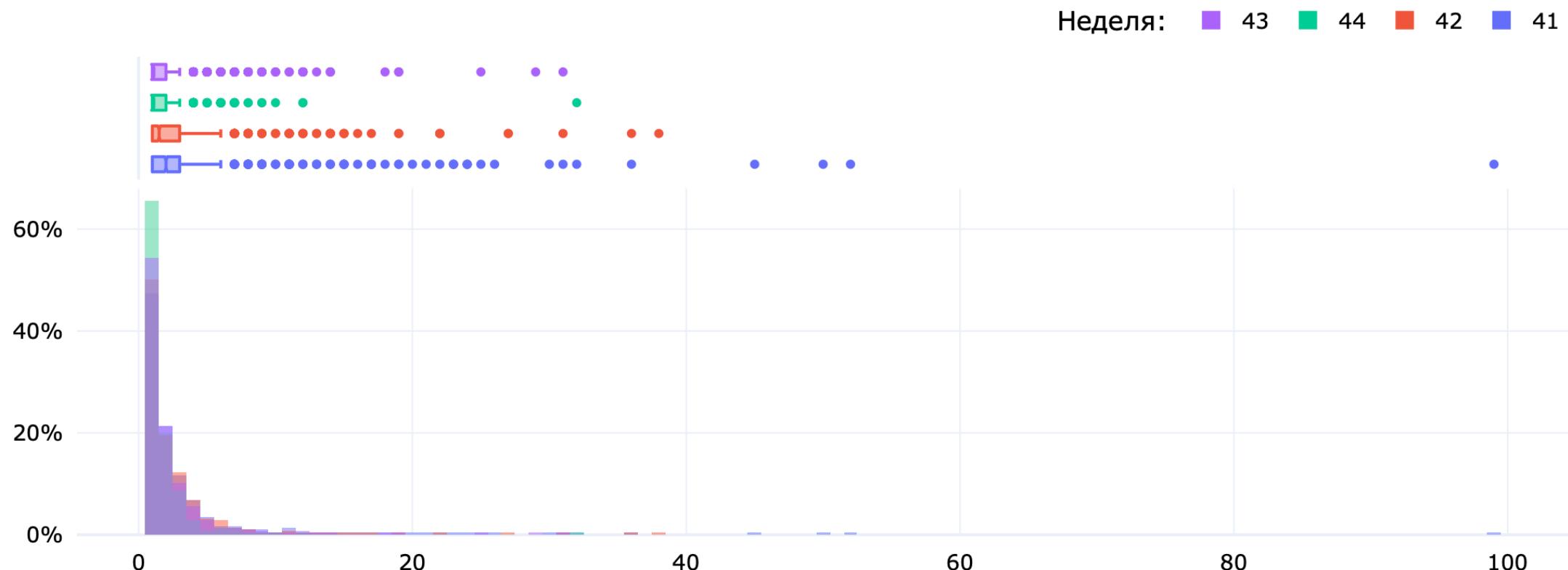
# Активность пользователей на протяжении дня



Главная активность в приложении начинается примерно с 8 утра и идет на убыль после 22 вечера.

Наблюдается спад активности с 17 по 19. Возможно, в это время люди едут с работы домой.

# Количество сессий на пользователя

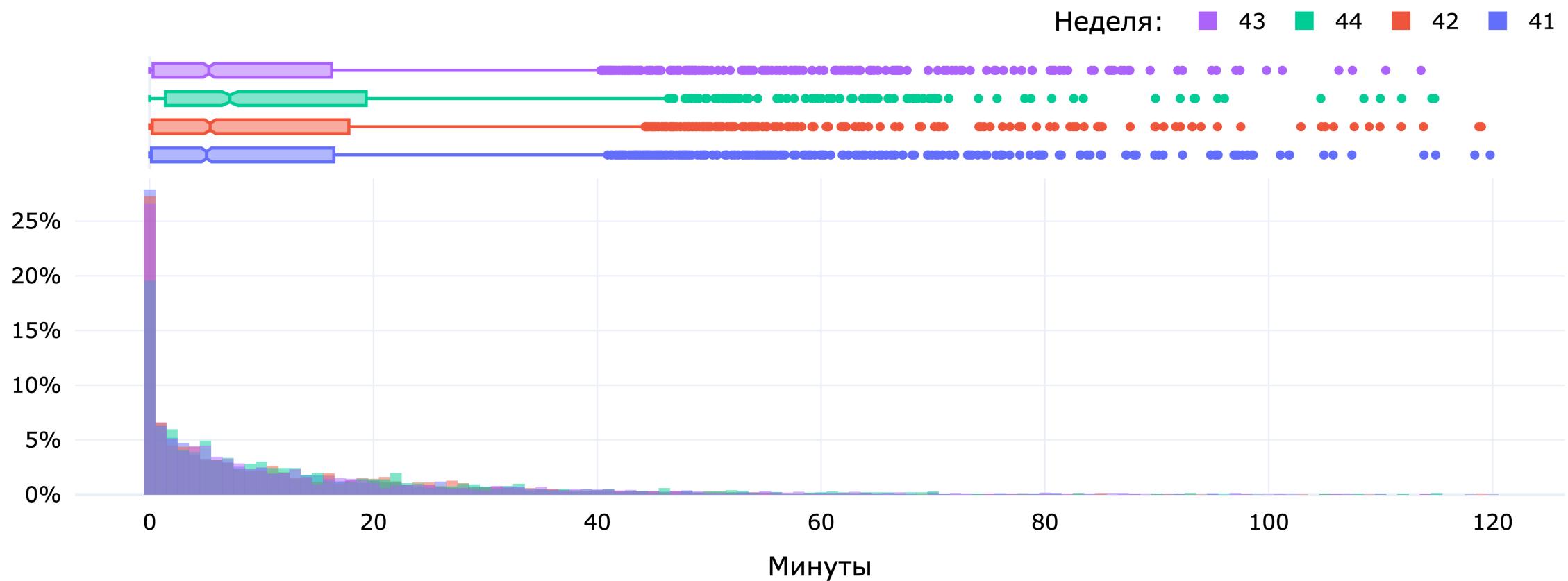


У половины пользователей была только 1 сессия.

У 20% пользователей было 2 сессии, и у 10% – 3.

У когорт 44 недели самая высокая доля пользователей только с 1 сессией. На 10% выше, чем у остальных когорт.

# Продолжительность сессий



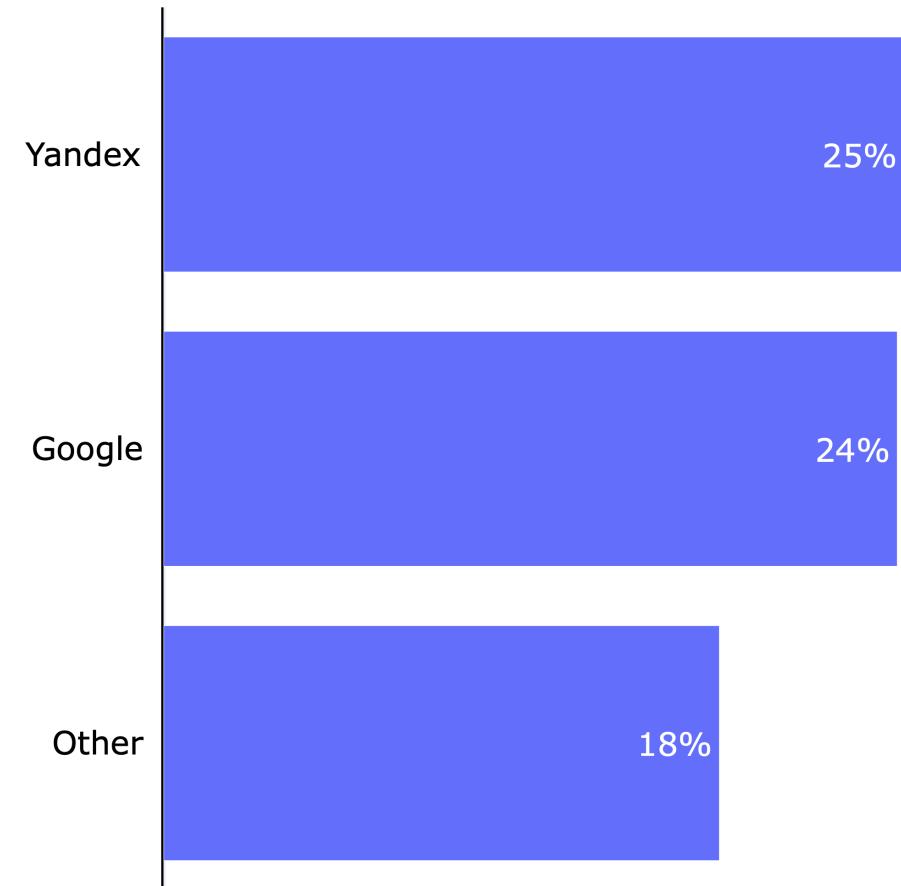
Продолжительность сессий имеет экспоненциальное распределение. Больше всего сессий, которые длились меньше минуты.

Медианная продолжительность сессии первых когорт составляет ~5 минут, однако у 44 когорты она больше 7 минут.

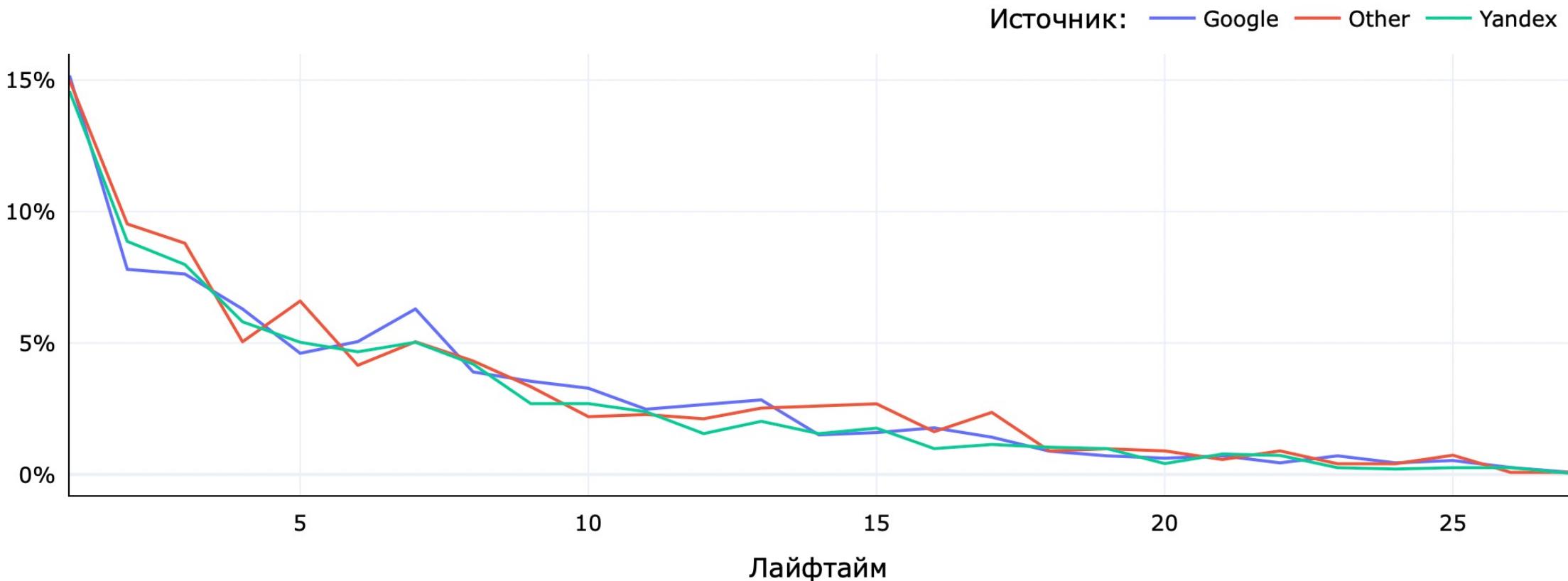
## Конверсия пользователей в просмотр контактов по источнику

У Яндекса и у Google примерно 24% пользователей просматривают контакты.

У пользователей с других источников конверсия всего 18%.



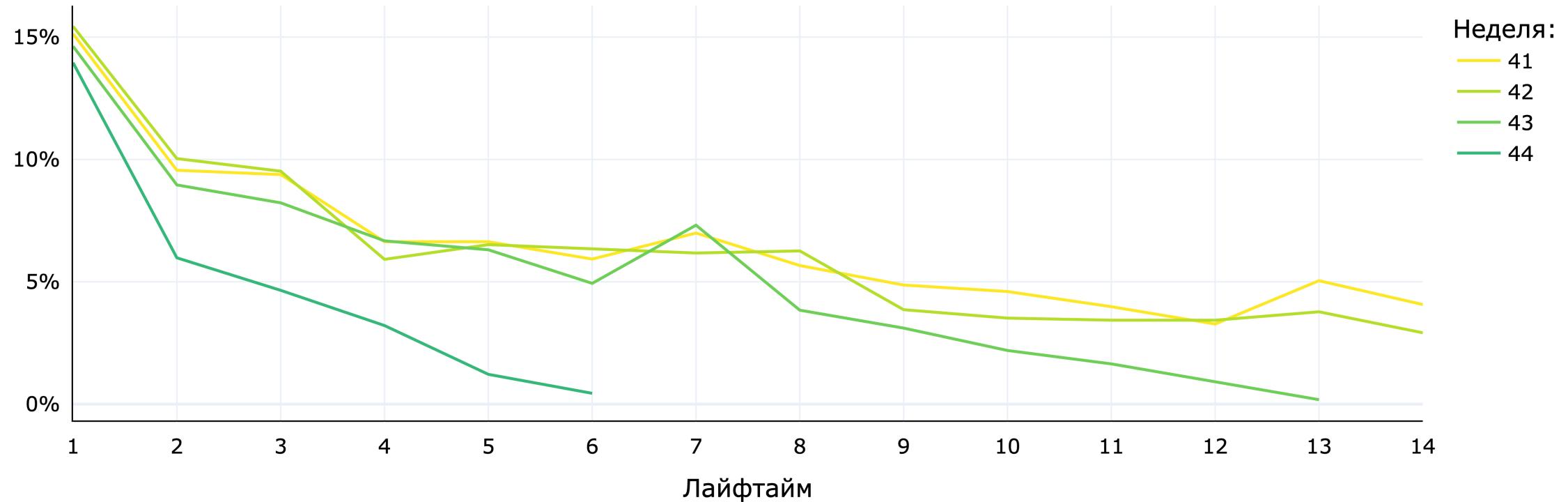
# Удержание пользователей



Источники не отличаются по удержанию пользователей.

Меньше 5% пользователей зашли в приложение через неделю.

## Удержание пользователей по когортам



У когорты 44 недели гораздо ниже  
удержание, чем у первых 3 недель.

У когорты 43 недели удержание на 13  
лайфтайм сильно ниже, чем у когорт  
41 и 42 недель.

## Возможные причины уменьшения удержания:

- **Изменилось распределение событий**

Например, рекомендованные объявления стали хуже совпадать с интересами пользователей. Вероятно, среднее количество пролистанных объявлений уменьшится.

- **Технические проблемы в приложении**

Количество пользователей растет, нагрузка на сервера увеличивается, приложение начинает дольше работать. Можно ожидать увеличение среднего времени между событиями в сессиях.

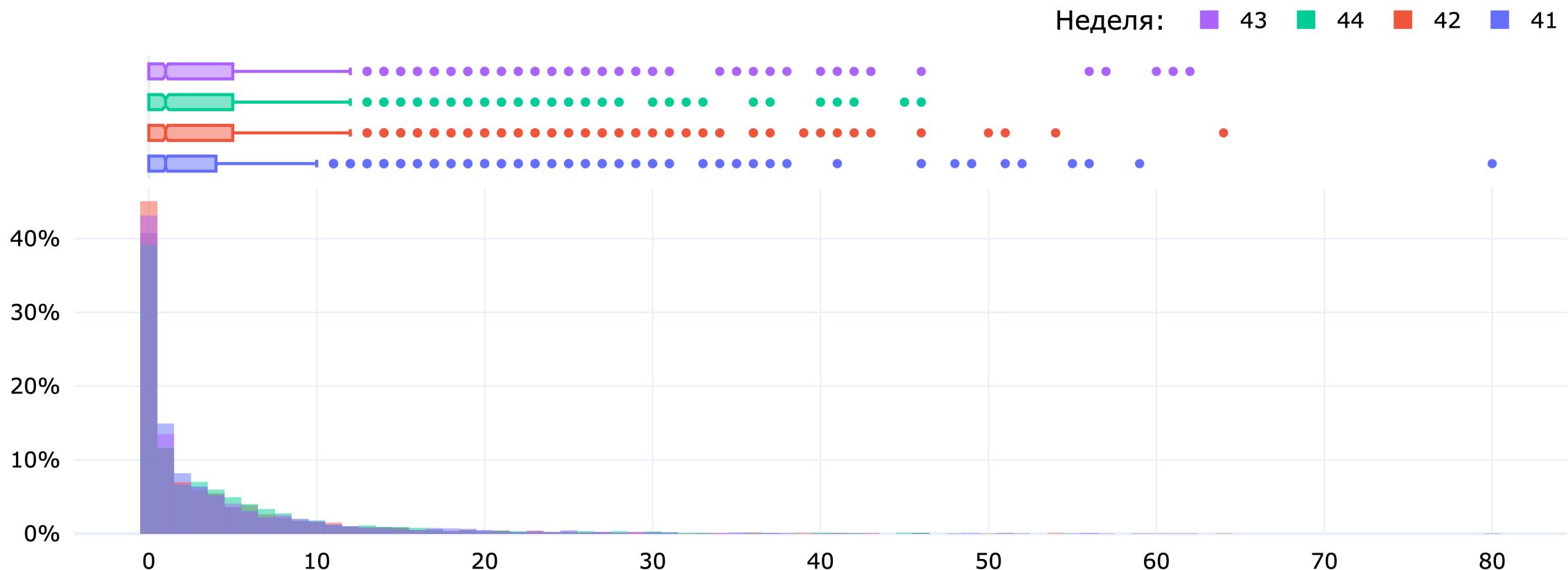
- **Приложение улучшилось**

Пользователи стали быстрее находить то, что им нужно. Вероятно, увеличилась бы доля просмотра контактов и фотографий.

- **Внешние факторы**

Приложение стало меньше рекламироваться и популярность упала.

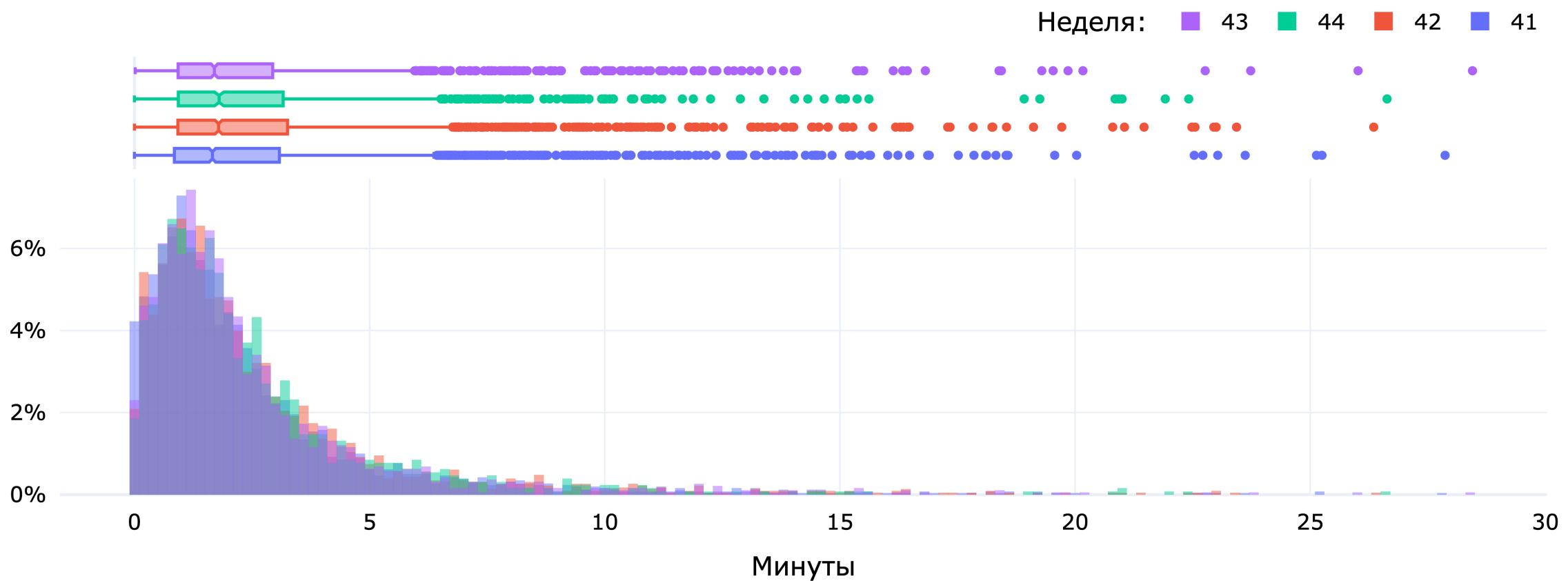
## Количество событий tips\_show на сессию



Когорты не отличаются по количеству пролистанных объявлений.

Значит, приложение вряд ли как-то изменилось: стало лучше или хуже.

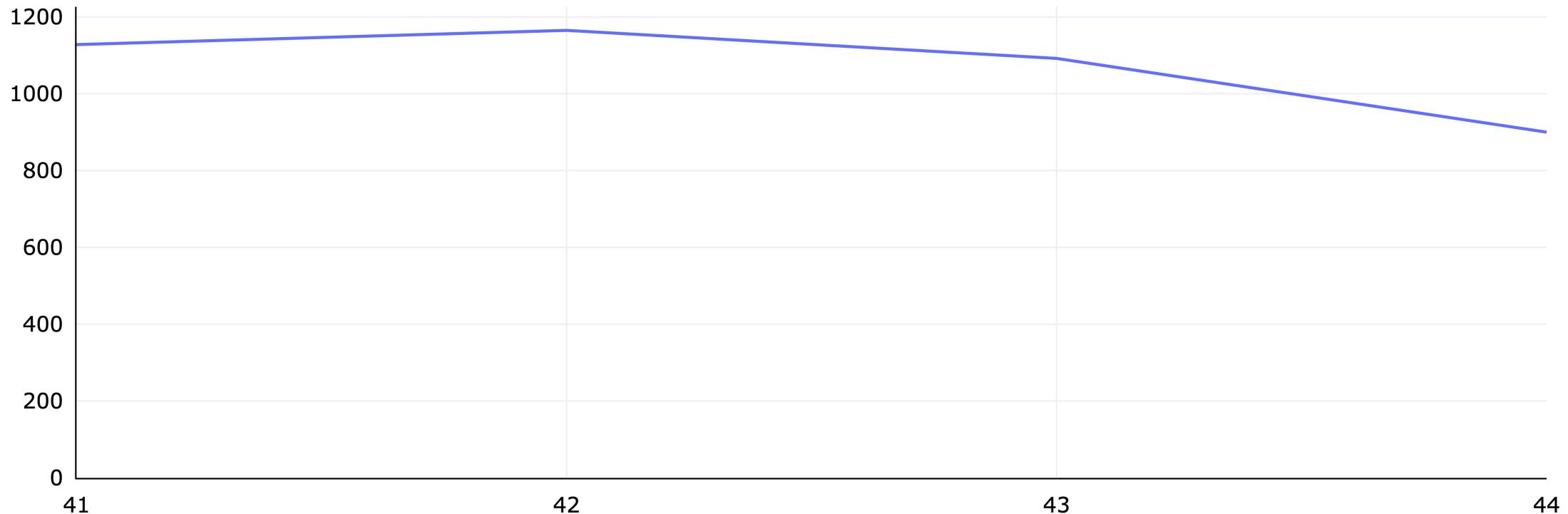
# Среднее время между событиями в сессиях



Медианное время между событиями  
составляет ~1.5 минуты.

Когорты практически не  
отличаются, значит, скорее  
всего, приложение не тормозит.

## Количество новых пользователей в неделю

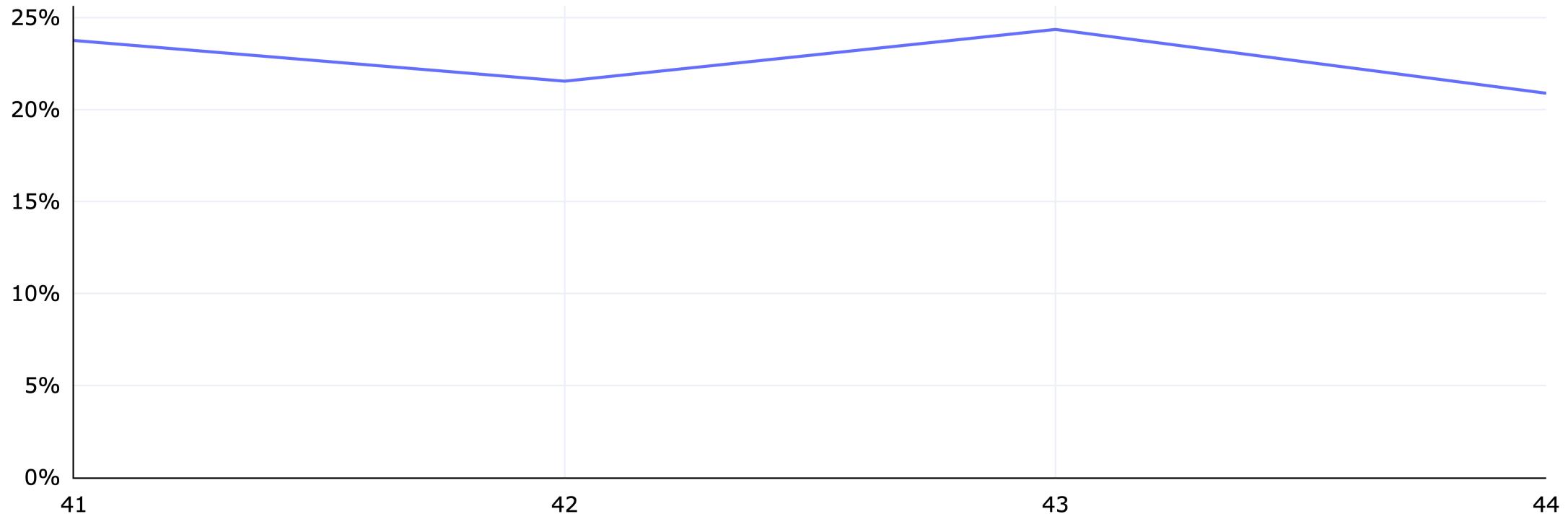


В среднем в неделю приходит 1000 новых пользователей.

После 42 недели количество новых пользователей в неделю начало немного снижаться.

Это может говорить о падении популярности приложения.

## Конверсия пользователей в просмотр контактов по неделям



В среднем 20-25% пользователей просматривают контакты.

За 44 неделю самая низкая конверсия в просмотре контактов – на 3-4% меньше, чем у 41 и 43 недель.

# Итоги по удержанию пользователей

Удержание не зависит от источника.

При разбиении пользователей на недельные когорты, образовалось 4 когорты: с 41 недели по 44. Удержание у пользователей, пришедших в 44 неделю, гораздо ниже, чем у остальных когорт.

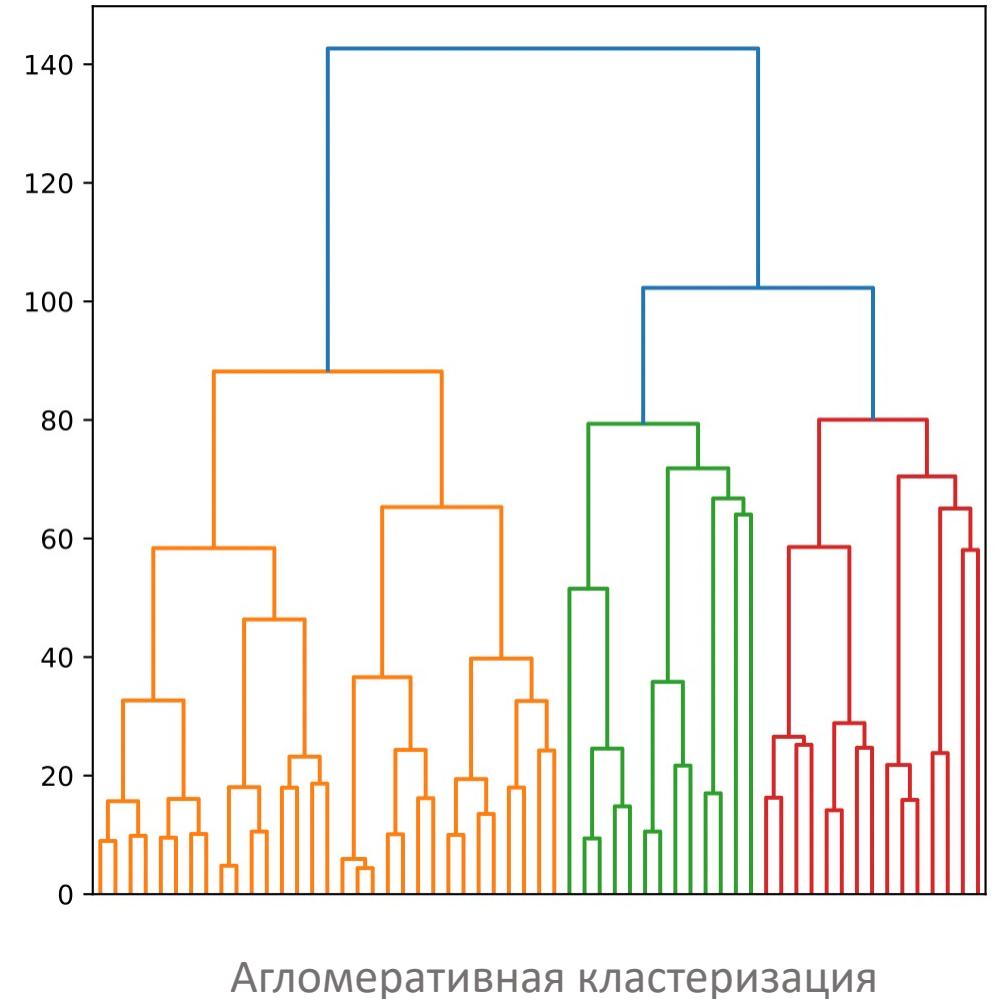
- Когорты не отличаются в распределениях событий на сессию. Распределение просмотренных фотографий и контактов не изменилось, что говорит о том, что удержание упало не из-за того, что приложение как-то изменилось.
- Когорты не отличаются по промежутку времени между событиями в сессиях. Это значит, что приложение не стало зависать или слишком медленно работать. Значит, падение удержания не связано с техническими проблемами.
- Количество новых пользователей каждую неделю становится меньше, что может сигнализировать, что приложение становится менее популярным.

## Сегментация

Агломеративная кластеризация выделила 3 основных кластера.

Эксперименты с методом k-средних и разным количеством кластеров показали, что самый высокий коэффициент силуэта достигается при 2-х кластерах.

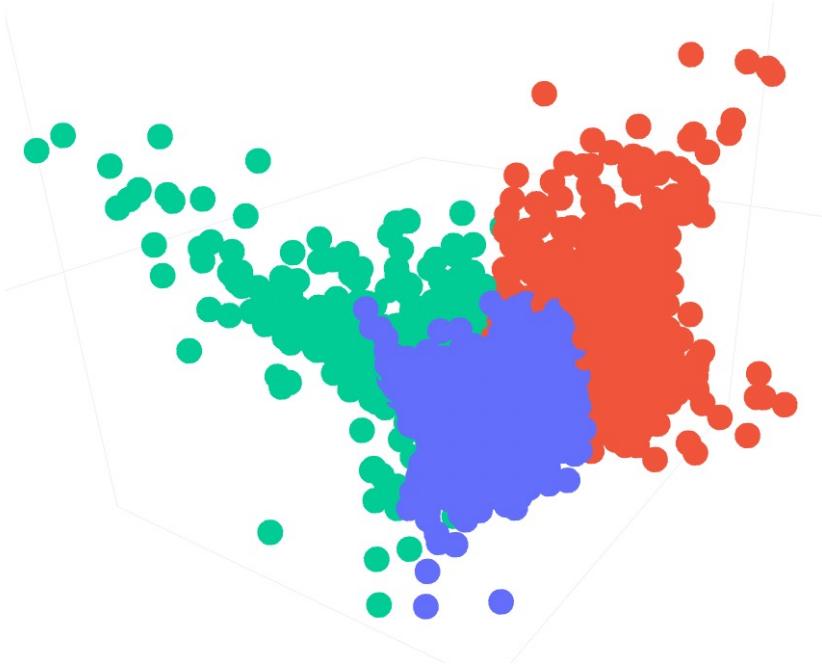
Однако было выбрано 3 кластера, так как 2 кластера могут недостаточно сегментировать пользователей.



Агломеративная кластеризация

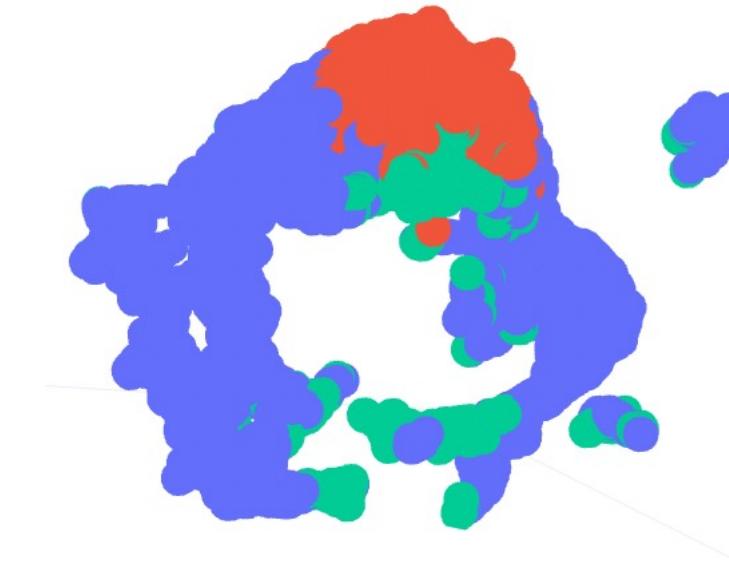
## Проекции кластеров в 3-мерное пространство

PCA



Кластеры очень близко расположены друг к другу, и сложно сказать, что они разделимы.

UMAP

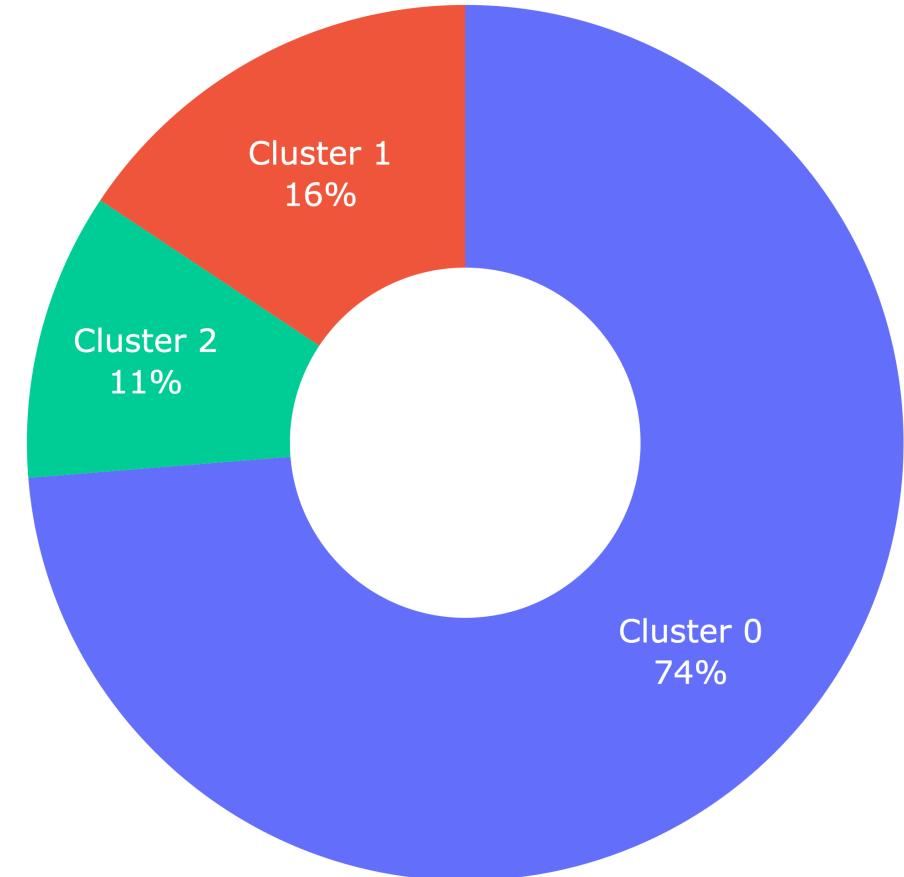


Вероятно, данные сами по себе не обладают четкой структурой.

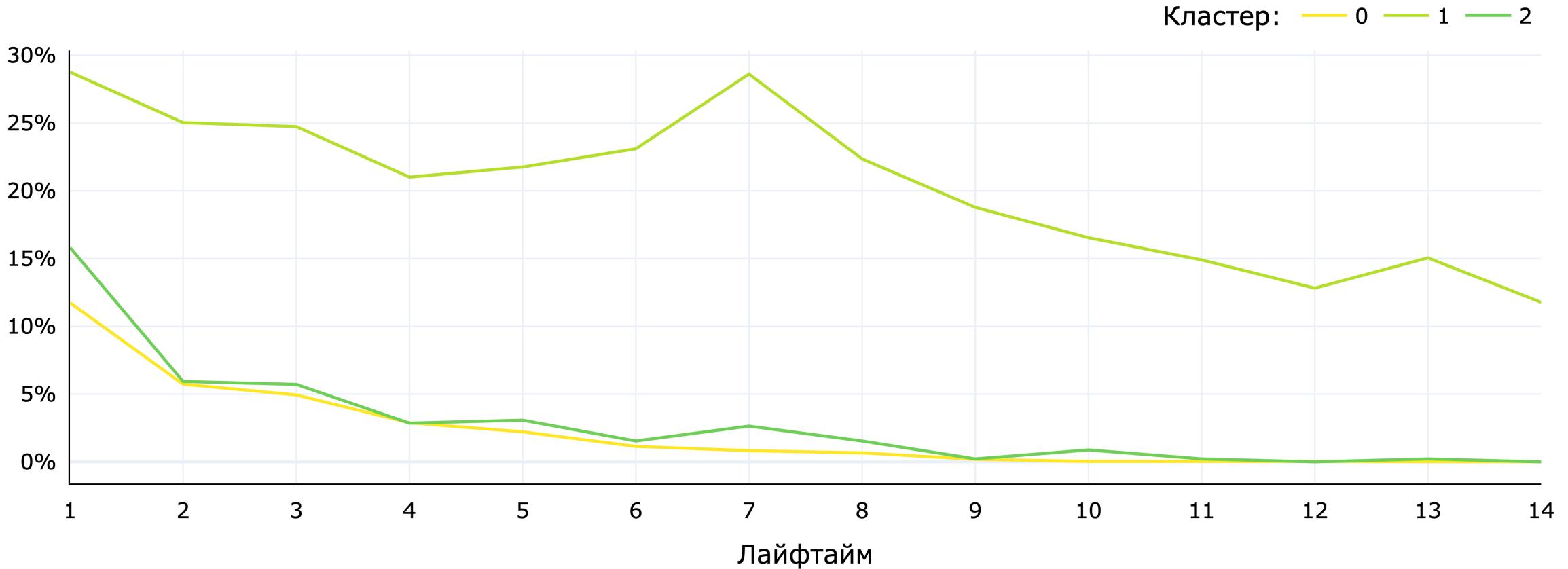
## Размеры кластеров

В кластерах присутствует большой дисбаланс по количеству пользователей.

На кластер 0 приходится ~74% всех пользователей, а на 2-й кластер только ~11%.



## Удержание по кластерам



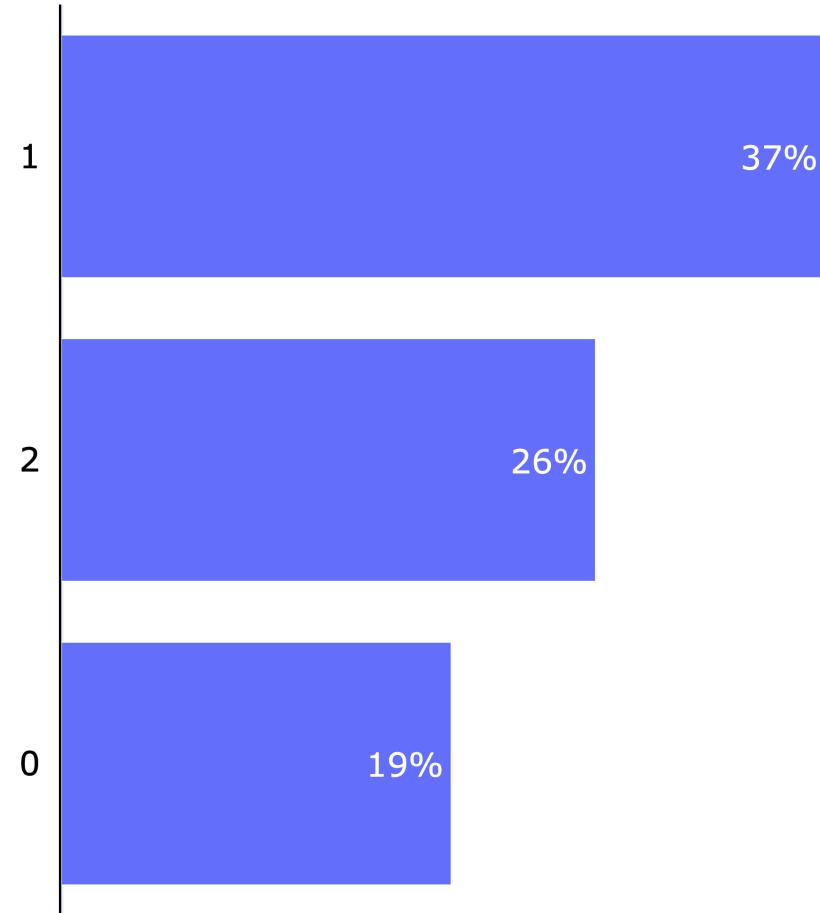
У кластера 1 гораздо выше удержание.  
Вероятно, это кластер с самыми  
лояльными пользователями.

Кластеры 0 и 2 практически не  
отличаются по удержанию, хотя у  
кластера 2 она чуть выше.

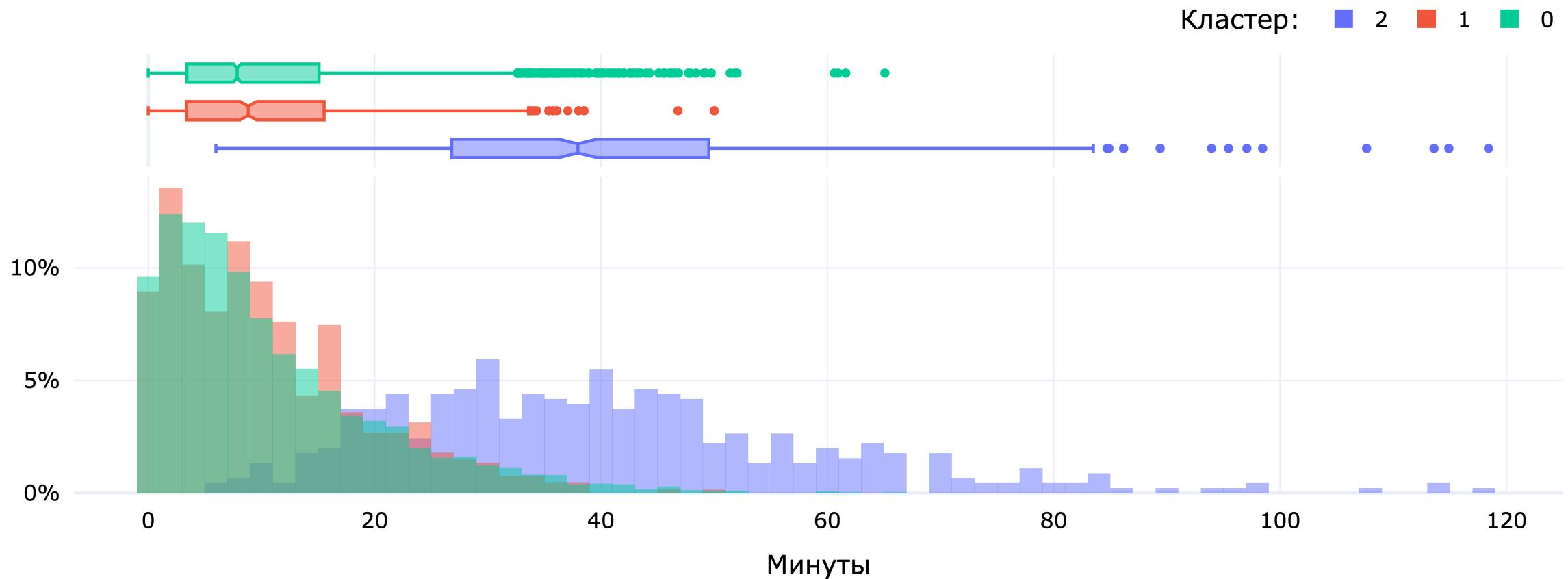
## Конверсия пользователей по кластерам

У кластера 1 выше конверсия, чем у остальных кластеров.

У 2-го кластера немного выше конверсия, чем у 0-го кластера.



# Продолжительность сессий по кластерам



У кластеров 0 и 1 распределения экспоненциальные и совпадают, а вот у кластера 2 распределение похоже на нормальное и сильно правее.

Пользователи 2-го кластера проводят в приложении в среднем по 40 минут, а пользователи первых двух кластеров – по 8-9.

## Итоги по сегментации пользователей

Пользователи сегментируются на 3 кластера:

- Кластер 0 представляет большинство пользователей – 74% всех пользователей. В этом кластере ниже всего удержание и конверсия по сравнению с остальными кластерами.
- Кластер 1 представляет самых лояльных пользователей – 16% всех пользователей. Здесь самые высокие удержание и конверсия в просмотр контактов.
- Кластер 2 представляет пользователей, которые больше всего проводят времени в приложении – 10%. Здесь такое же удержание, как и у большинства пользователей, однако выше конверсия в просмотр контактов.

## Проверка гипотез

- Гипотезы для конверсии по источникам (где А и В – пара источников):

**Нулевая гипотеза:** между источниками А и В нет разницы в конверсии в просмотр контактов.

**Альтернативная гипотеза:** между источниками А и В есть разница в конверсии в просмотр контактов.

- Гипотезы для конверсии по кластерам (где А и В – пара кластеров):

**Нулевая гипотеза:** между кластерами А и В нет разницы в конверсии в просмотр контактов.

**Альтернативная гипотеза:** между кластерами А и В есть разница в конверсии в просмотр контактов.

## Итоги проверки гипотез

Было проверено 5 гипотез с помощью t-теста. Чтобы получить групповую вероятность ошибки первого рода меньше 0.05, к уровню значимости была применена поправка Шидака, что дало уровень значимости 0.01 для каждой гипотезы.

- Конверсии в просмотр контактов различаются между источником Google/Яндекс и остальными источниками. Между самими источниками Google и Яндекс нет статистически значимой разницы в конверсии.
- Между всеми кластерами есть статистически значимая разница в конверсии в просмотр контактов.

## Краткие итоги

- Приложение может терять популярность.
- Кластер с 16% пользователей имеет самое высокое удержание и конверсию в просмотр контактов.
- Между всеми кластерами есть статистически значимая разница в конверсии в просмотр контактов.
- Конверсии в просмотр контактов различаются между источником Google/Яндекс и остальными источниками. Между самими источниками Google и Яндекс нет статистически значимой разницы в конверсии.