



# Исследование банковских клиентов

Исследование выполнил Александр Глебовский  
в рамках курса по анализу данных

# Исследование банковских клиентов

Данные взяты с сайта kaggle (<https://www.kaggle.com/>)

Охвачен период 2016/08/02 - 2016/09/16

Количество строк в исходном материале = число транзакций = 985322

## Результаты исследования

1. Уникальных клиентов в рассматриваемом периоде: - 839081
2. Общее количество транзакций: - 985322
3. Среднее количество транзакций на клиента, ( деление общего количества транзакций на число уникальных клиентов): - 1.17
4. Распределение и размах величин custaccountbalance (баланс клиента) и transactionamount(lnr) (сумма транзакций)

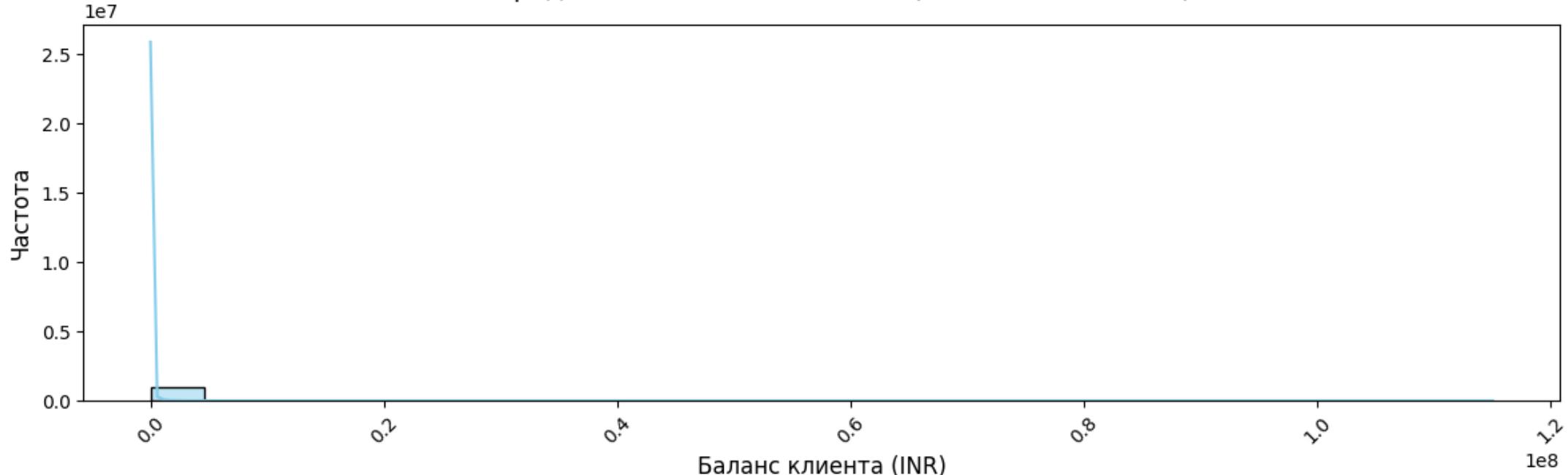
### Описательная статистика:

	custaccountbalance	transactionamount(lnr)
count	9.853220e+05	9.853220e+05 — количество наблюдений;
mean	1.060612e+05	1.452425e+03 — среднее;
std	8.179054e+05	6.139765e+03 — стандартное отклонение (разброс);
min	0.000000e+00	0.000000e+00 - min / max — границы;
25%	4.582132e+03	1.510000e+02 — квартили
50%	1.600630e+04	4.310000e+02 - медиана
75%	5.375908e+04	1.125000e+03 — квартили
max	1.150355e+08	1.560035e+06 - min / max — границы;

**Анализ гистограмм** \*\*Распределение баланса клиента (CustAccountBalance)\*\* —(верхний график.) и \*\*Распределение суммы транзакций (TransactionAmount (INR))\*\* —( нижний график.) показывает, что

- Оба распределения \*\*асимметричны с доминированием низких значений и редкими крупными выбросами\*\*. Это характерно для ситуации, где \*\*большая часть клиентов/операций — мелкие, а значительная масса средств/объёмов приходится на небольшую группу\*\*

### Распределение баланса клиента (custaccountbalance)



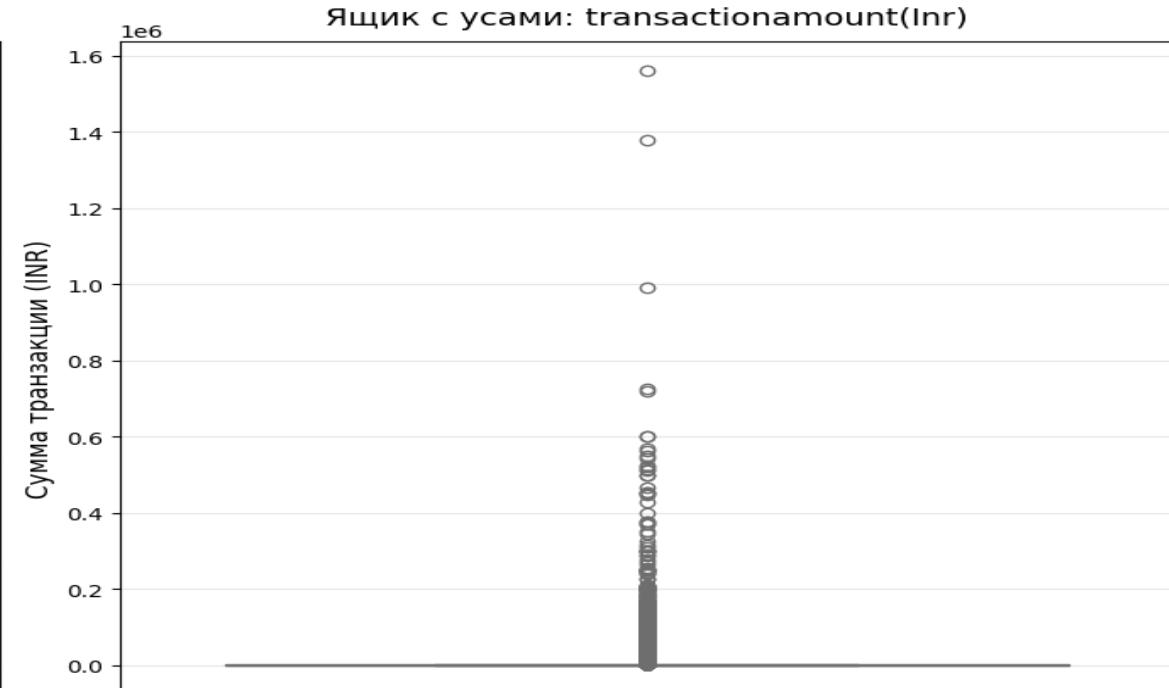
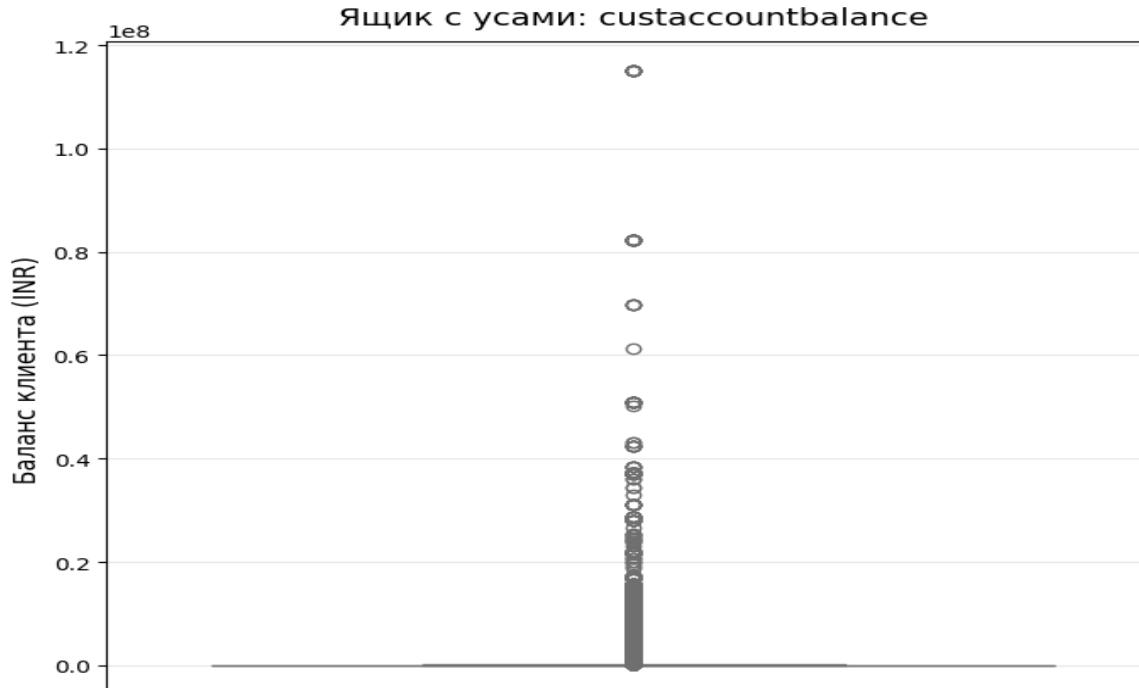
### Распределение суммы транзакций (transactionamount(INR))



## Анализ «ящика с усами» (box plot)

для показателей «CustAccountBalance» и «TransactionAmount (INR)» показывает, что

- **Размах значений:** - баланс имеет гораздо более широкий диапазон (до 100 млн INR), чем сумма транзакции (до 1 млн INR) — логично, так как баланс аккумулирует средства, а транзакция фиксирует единичное движение.
- **Асимметрия:** - оба распределения асимметричны, но у баланса асимметрия более выражена (большее влияния крупных значений).
- **Выбросы:** - в обоих случаях присутствуют, но у баланса они значительно крупнее (на порядки выше, чем у транзакций).
- **Концентрация в нуле:** - оба показателя демонстрируют наличие нулевых значений, что требует дополнительного анализа.



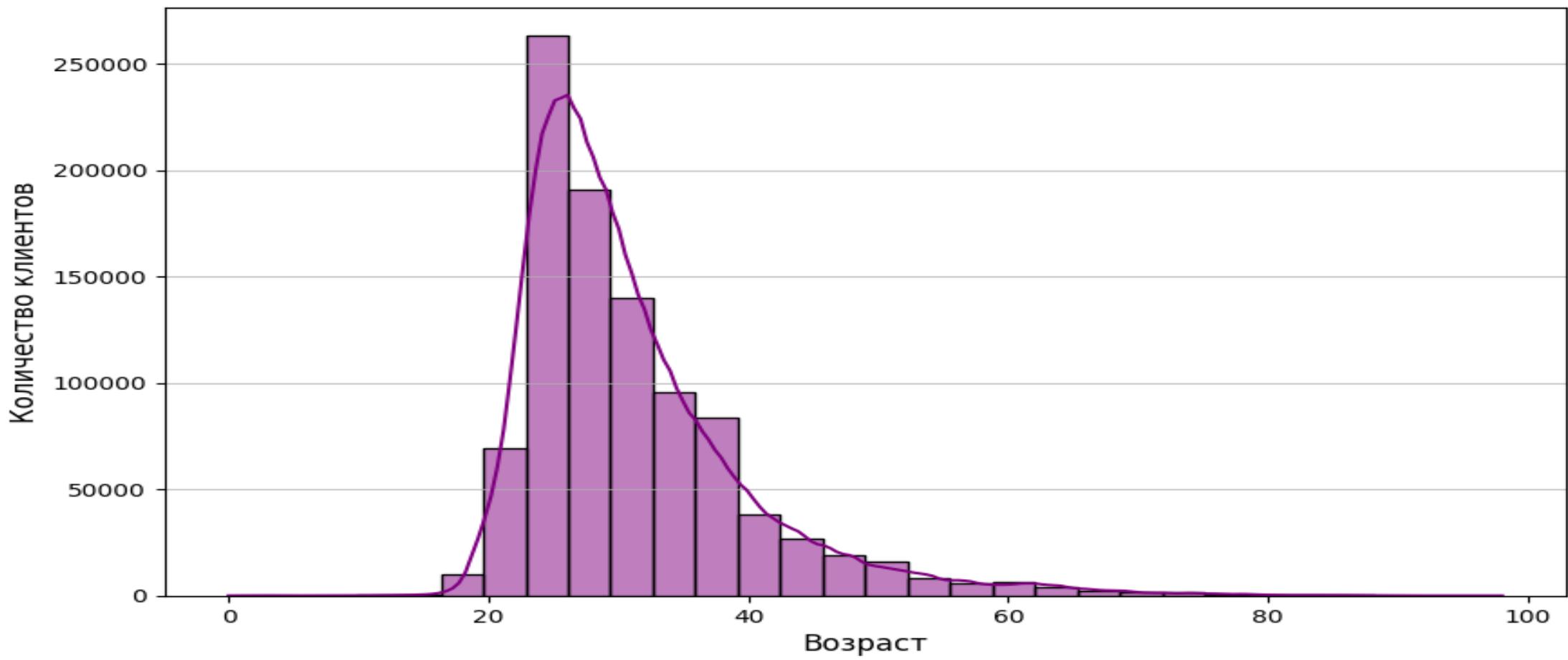
## 5. Возраст клиентов.

Описательная статистика возраста клиентов:

count 985322.000000 - количество наблюдений;  
mean 31.029537 - среднее  
std 8.757113 - стандартное отклонение (разброс)  
min 0.000000- min / max — границы  
25% 25.000000— квартили  
50% 29.000000 - медиана  
75% 34.000000 — квартили  
max 98.000000 - min / max — границы;

На графике распределения возраст наибольшего числа клиентов - 25 лет.

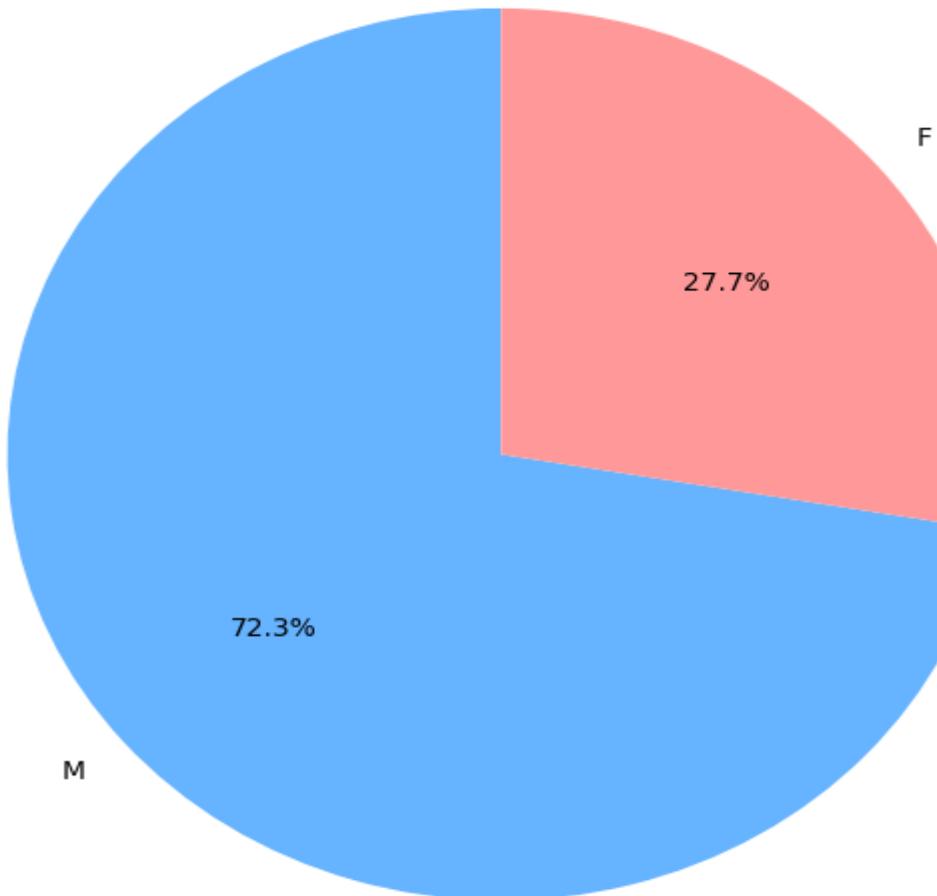
## Распределение возраста клиентов



возраст наибольшего числа клиентов - 25 лет.

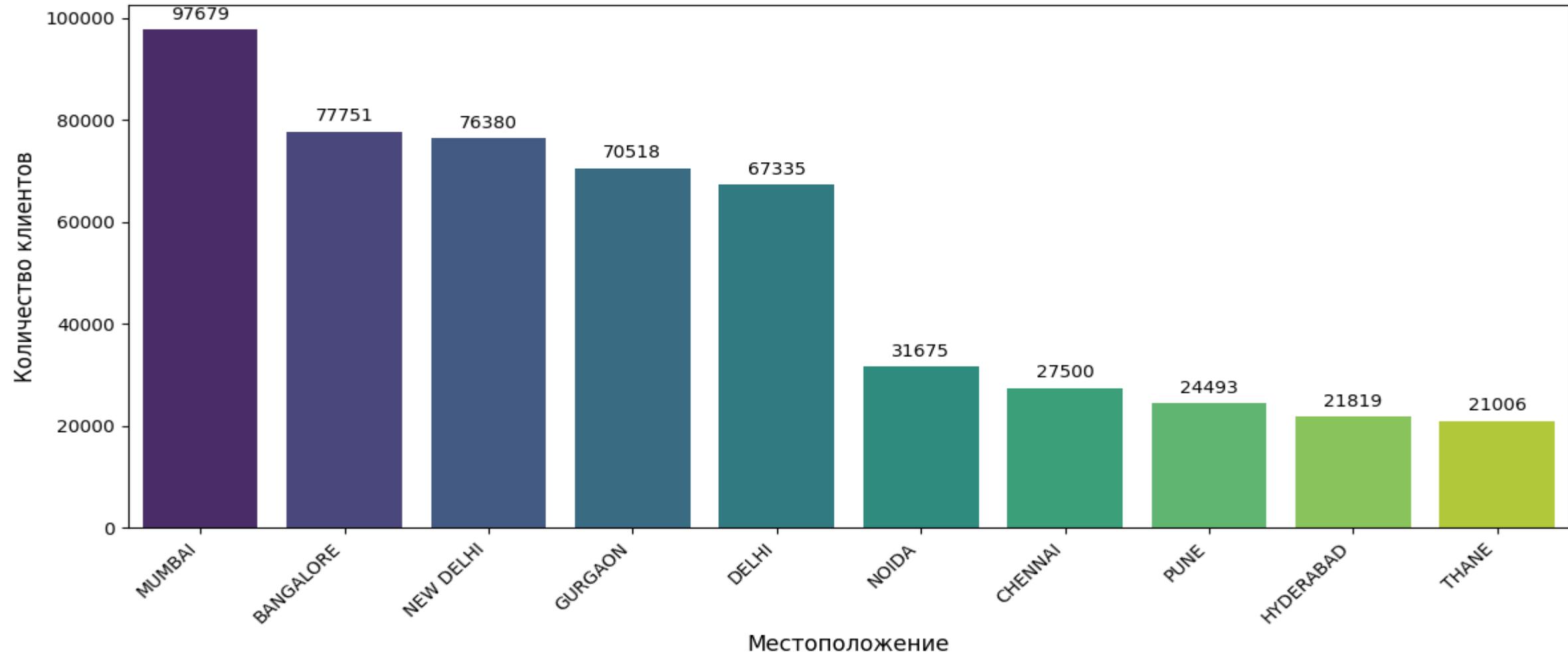
6. Половой состав клиентов: мужчины - 72,3%, женщины - 27,7%

Распределение клиентов по полу



7. Анализ места жительства клиентов показывает первая тройка по кол-ву клиентов: Мумбай, Бангалор, Нью Дели.

Топ 10 мест жительства клиентов по количеству



## 8. Динамика количества и суммы транзакций по датам,:

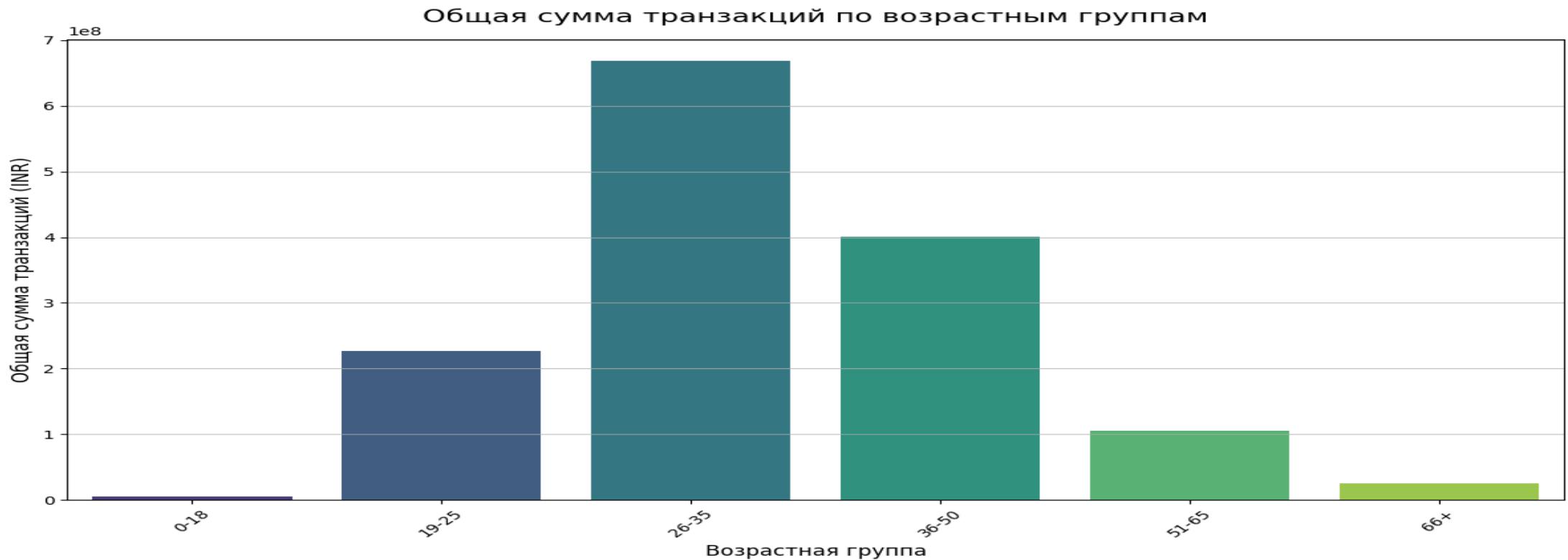
Весь август и половину сентября кол-во транзакций держится примерно на одном уровне 20000, а к середине октября падает почти до нуля.

## 9. Динамика суммы транзакций ведет себя аналогично



10. Согласно результатам анализа, наиболее платежеспособной является возрастная группа «26–35 лет».

- \* Клиенты были успешно разделены на шесть возрастных групп: «0–18 лет», «19–25 лет», «26–35 лет», «36–50 лет», «51–65 лет» и «66+ лет».
- \* Возрастная группа «от 26 до 35 лет» была самой многочисленной среди всех категорий.
- \* Возрастная группа «от 26 до 35 лет» также совершила наибольшее количество транзакций, что делает её самой платежеспособной группой, за которой следует возрастная группа «от 36 до 50 лет».
- \* На гистограмме наглядно показаны общие суммы транзакций в этих возрастных группах, что ясно демонстрирует доминирование группы «от 26 до 35 лет».



11. Сравнение 10 крупнейших городов по количеству транзакций и общей сумме транзакций показывает следующее:

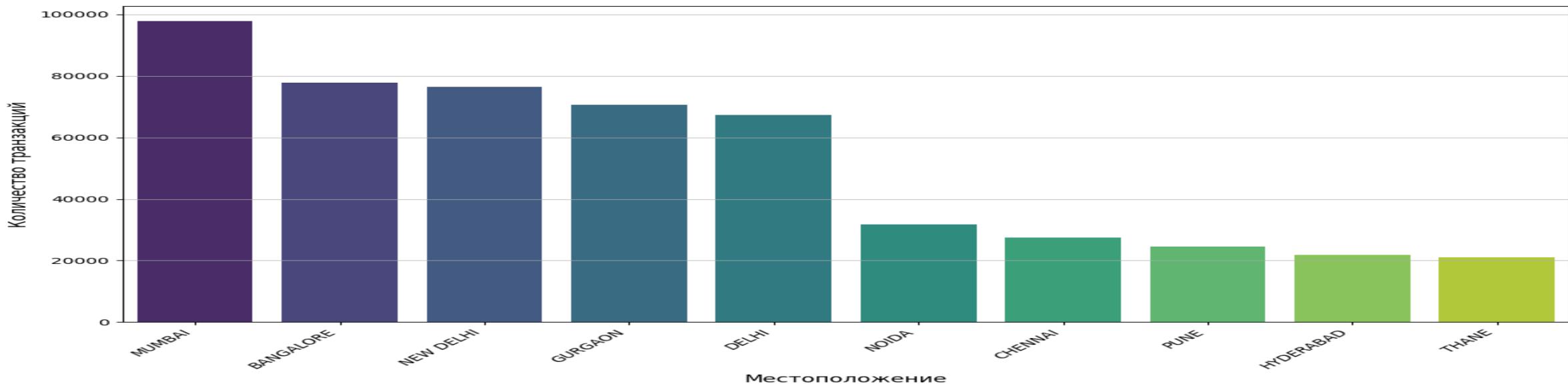
- \* \*\*Общие города\*\*: в обоих списках встречаются девять общих городов: «Мумбаи», «Бангалор», «Нью-Дели», «Гургаон», «Дели», «Ноида», «Ченнаи», «Пуна» и «Хайдарабад». Это указывает на то, что в этих крупных городских центрах наблюдается стабильно высокая активность и стоимость транзакций.
- \* \*\*Уникальность по количеству транзакций\*\*: «Тане» входит в топ-10 по количеству транзакций. Это говорит о том, что в Тане большой объём транзакций, но их стоимость относительно невысока, что приводит к снижению общей стоимости транзакций.
- \* \*\*Уникальность по общей сумме транзакций\*\*: «Колката» входит в топ-10 по общей сумме транзакций. Это означает, что, хотя в Калькутте может быть меньше транзакций, чем в таких местах, как Тейн, сумма транзакций в Калькутте значительно выше, что позволяет ей входить в топ-10.

### Основные выводы \* \*\*Сильное пересечение\*\*: большое количество общих мест (9 из 10) указывает на то, что самые густонаселённые и экономически активные города, как правило, лидируют как по количеству, так и по общей сумме транзакций.

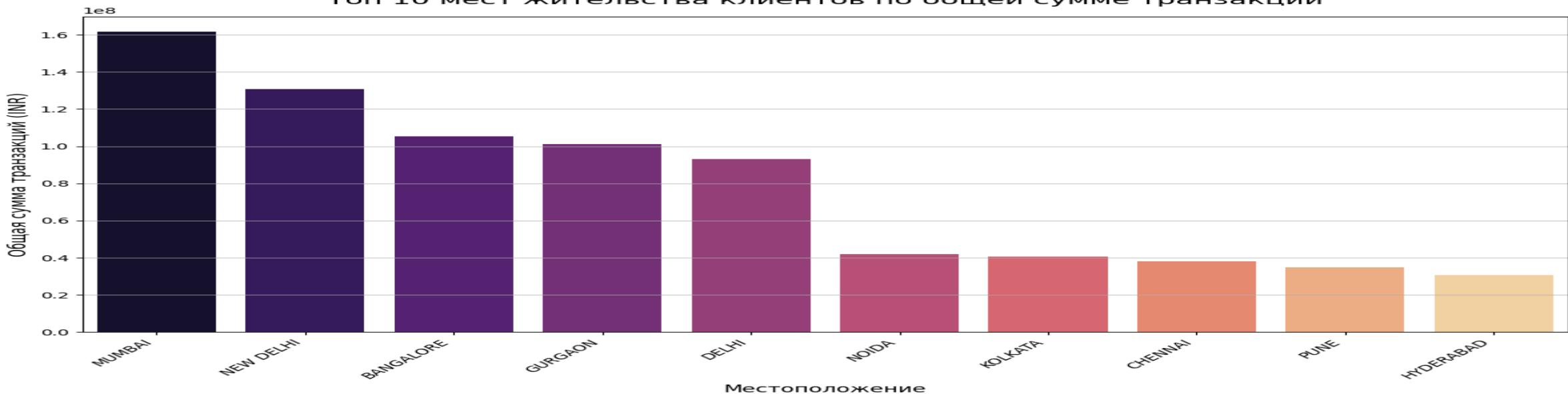
\* \*\*Разница в объеме и стоимости\*\*: Тане — пример района, где объем транзакций \*выше\*, а их стоимость относительно ниже. С другой стороны, Калькутта — пример района, где более высокая \*стоимость\* транзакции вносит значительный вклад в общую сумму, даже если количество транзакций не так велико.

\* \*\*Доминирование Мумбаи\*\*: Мумбаи неизменно занимает высокие позиции в обеих категориях, подтверждая свой статус основного финансового центра с высоким объемом и стоимостью транзакций.

Топ 10 мест жительства клиентов по количеству транзакций

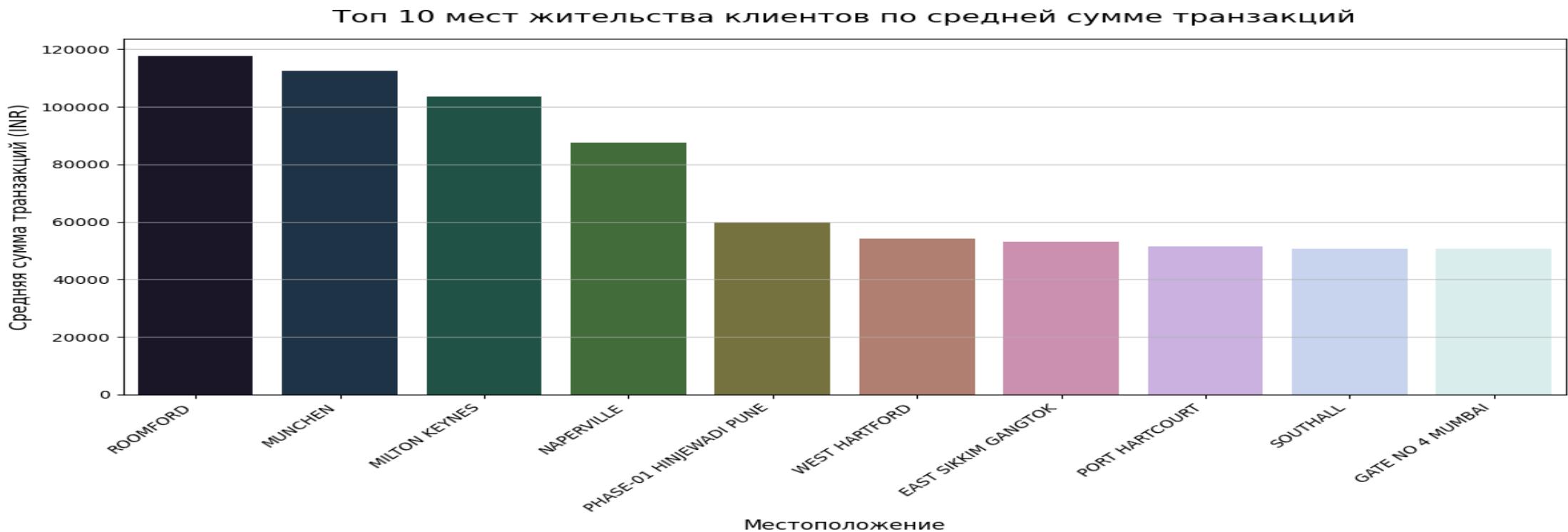


Топ 10 мест жительства клиентов по общей сумме транзакций



### Анализ средних сумм транзакций в зависимости от местоположения выявляет интересные закономерности в сравнении с количеством транзакций и их общей суммой:

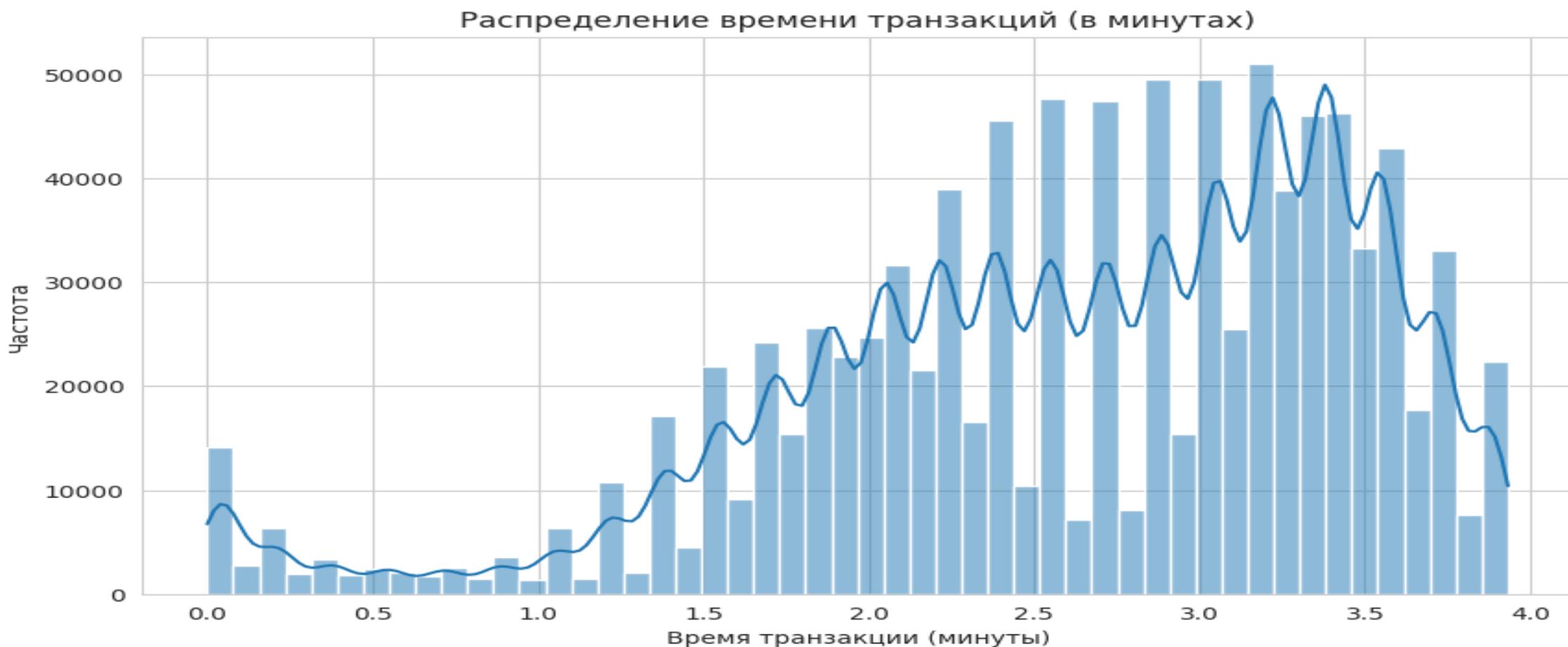
- \* В топ-10 локаций по средней сумме транзакций входит ряд локаций, для которых в первую очередь характерны очень высокие суммы отдельных транзакций (например, «ROOMFORD» — 117 621 индийских рупий, «MUNCHEN» — 112 592 индийских рупий).
- \* Эти локации сильно отличаются от топ-10 локаций по количеству транзакций и общей сумме транзакций, среди которых преобладают крупные индийские города.
- \* Это говорит о том, что в местах с большим количеством транзакций и их общей суммой не обязательно самые высокие средние значения транзакций, и наоборот.



## 12. распределение времени транзакций

График времени транзакций показывает: время транзакций варьируется от минимальных значений (около 0 минут) до максимальных (около 4 минут).

Общая тенденция: большинство транзакций сосредоточено в нижней части диапазона (до 2–3 минут), что указывает на относительно быстрые операции.



----Анализ ящика с усами показывает:

Основная масса транзакций (50%) укладывается в диапазон от ~2 до ~3,3 минут (границы ящика).

Медиана (Q2) находится около 2,7–2,8 минут — это среднее время транзакции по выборке.

Первый quartиль (Q1) — около 2 минут (25% транзакций выполняются быстрее этого времени).

Третий quartиль (Q3) — около 3,3 минут (75% транзакций укладываются в это время).

Усы (диапазон «нормальных» значений):

Нижний ус: начинается около 0 минут — минимальное «нормальное» время транзакции.

Верхний ус: заканчивается около 4 минут — максимальное «нормальное» время транзакции.

Это говорит о том, что 95% транзакций укладываются в диапазон 0–4 минут.

Выбросы:

На графике видны выбросы слева (несколько точек около 0 минут) — аномально быстрые транзакции.

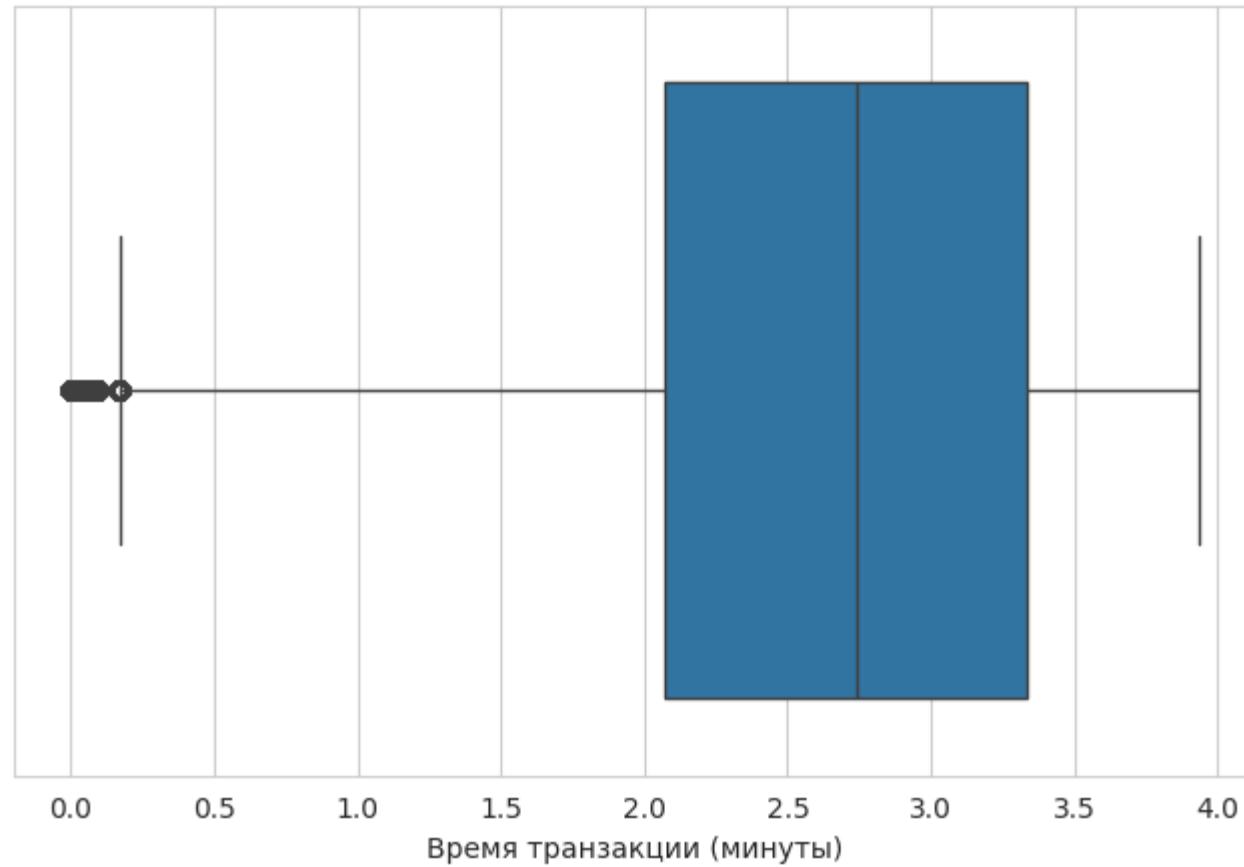
Выбросов справа (длительных транзакций) не наблюдается, что говорит о хорошем контроле времени обработки.

Асимметрия:

Bxplot скошен влево (длинный ус слева, короткий — справа), что указывает на наличие аномально быстрых транзакций (близких к 0 минут).

Основная масса транзакций сконцентрирована в правой части распределения (ближе к медиане и Q3).

Boxplot времени транзакций



Ключевые выводы:

Типичное время транзакции: 2,7–3,3 минуты (интерквартильный размах).

Минимальное время: около 0 минут (с учётом выбросов).

Максимальное «нормальное» время: около 4 минут.

Аномалии: присутствуют очень быстрые транзакции (выбросы слева), которые требуют дополнительного анализа.

Отсутствие длительных выбросов: это положительный сигнал — система обработки транзакций работает стабильно, без значительных задержек.

---- Анализ графика «Плотность распределения времени транзакций»

Пик плотности (мода): основной пик наблюдается около 3–3,3 минут. Это означает, что наиболее часто встречающееся время транзакции — около 3 минут.

Форма распределения: распределение асимметрично (скошено влево). Основная масса транзакций сосредоточена в правой части графика (ближе к 3 минутам), а левая часть (ближе к 0) имеет более низкие значения плотности.

Волнообразные колебания: небольшие пики и спады в диапазоне 1–3 минут указывают на многообразие типов транзакций с разным типичным временем обработки.

### Плотность распределения времени транзакций





**Спасибо за внимание!**

Исследование выполнил Александр Глебовский  
E-mail: vizitor1@mail.ru