ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Московский институт электроники и математики им. А.Н.Тихонова национального исследовательского университета

«Высшая школа экономики»

# Отчёт по домашнему заданию

HW1 по курсу

«Методы и средства обработки больших данных»

**Студент**:

Дементьев К. А.

**Преподаватель**:  
 Доцент ДПМ МИЭМ НИУ ВШЭ

Вальба О. В.

**Москва 2025**

# Часть 1. Исследование данных и обработка данных для проведения последующей сегментации.

Представлены: данные по клиентам банка, которые имеют кредит

Размер данных: 10243 строк, 44 столбцов

Всего 44 переменных, две из которых, вариант и id, поскольку они явно не могут коррелировать с другими переменными.

Информация о типах данных:

19 категориальных

23 количественных

INCOME\_BASE\_TYPE кат

CREDIT\_PURPOSE кат

INSURANCE\_FLAG кат

DTI кол-во

SEX кат

FULL\_AGE\_CHILD\_NUMBER кол-во

DEPENDANT\_NUMBER кол-во

EDUCATION кат

EMPL\_TYPE кат

EMPL\_SIZE кат

BANKACCOUNT\_FLAG кат

Period\_at\_work кол-во

age кол-во

EMPL\_PROPERTY кат

EMPL\_FORM кат

FAMILY\_STATUS кат

max90days кол-во

max60days кол-во

max30days кол-во

max21days кол-во

max14days кол-во

avg\_num\_delay кол-во

if\_zalog кат

num\_AccountActive180 кол-во

num\_AccountActive90 кол-во

num\_AccountActive60 кол-во

Active\_to\_All\_prc кол-во

numAccountActiveAll кол-во

numAccountClosed кол-во

sum\_of\_paym\_months кол-во

all\_credits кол-во

Active\_not\_cc кол-во

own\_closed кол-во

min\_MnthAfterLoan кол-во

max\_MnthAfterLoan кол-во

dlq\_exist кат

thirty\_in\_a\_year кат

sixty\_in\_a\_year кат

ninety\_in\_a\_year кат

thirty\_vintage кат

sixty\_vintage кат

ninety\_vintage кат

Исследуем распределения по данным:

* Рассчитываем количество уникальных значений, нулевых и пустых значений + доля в % от общего количества;

================================================================================

АНАЛИЗ ПРОПУЩЕННЫХ И НУЛЕВЫХ ЗНАЧЕНИЙ

================================================================================

Анализ пропущенных и нулевых значений:

Тип данных Уникальных значений Пустых значений

INCOME\_BASE\_TYPE object 4 66

CREDIT\_PURPOSE object 10 0

INSURANCE\_FLAG float64 2 2

DTI object 60 129

SEX object 2 0

FULL\_AGE\_CHILD\_NUMBER int64 6 0

DEPENDANT\_NUMBER int64 3 0

EDUCATION object 9 0

EMPL\_TYPE object 9 8

EMPL\_SIZE object 7 127

BANKACCOUNT\_FLAG float64 3 2310

Period\_at\_work float64 357 2313

age float64 41 2310

EMPL\_PROPERTY object 12 2310

EMPL\_FORM object 6 6273

FAMILY\_STATUS object 6 6273

max90days float64 20 6335

max60days float64 19 6335

max30days float64 16 6335

max21days float64 15 6335

max14days float64 14 6335

avg\_num\_delay float64 1127 6601

if\_zalog float64 2 6585

num\_AccountActive180 float64 7 6585

num\_AccountActive90 float64 5 6585

num\_AccountActive60 float64 4 6585

Active\_to\_All\_prc float64 96 6585

numAccountActiveAll float64 13 6585

numAccountClosed float64 26 6585

sum\_of\_paym\_months float64 319 6585

all\_credits float64 30 6585

Active\_not\_cc float64 8 6585

own\_closed float64 10 6585

min\_MnthAfterLoan float64 97 6585

max\_MnthAfterLoan float64 134 6585

dlq\_exist float64 2 6585

thirty\_in\_a\_year float64 2 6585

sixty\_in\_a\_year float64 2 6585

ninety\_in\_a\_year float64 2 6585

thirty\_vintage float64 2 6585

sixty\_vintage float64 2 6585

ninety\_vintage float64 2 6585

Доля пустых (%) Нулевых значений Доля нулевых (%)

INCOME\_BASE\_TYPE 0.64 0 0.00

CREDIT\_PURPOSE 0.00 0 0.00

INSURANCE\_FLAG 0.02 3940 38.47

DTI 1.26 0 0.00

SEX 0.00 0 0.00

FULL\_AGE\_CHILD\_NUMBER 0.00 6044 59.01

DEPENDANT\_NUMBER 0.00 10212 99.70

EDUCATION 0.00 0 0.00

EMPL\_TYPE 0.08 0 0.00

EMPL\_SIZE 1.24 0 0.00

BANKACCOUNT\_FLAG 22.55 6326 61.76

Period\_at\_work 22.58 0 0.00

age 22.55 0 0.00

EMPL\_PROPERTY 22.55 0 0.00

EMPL\_FORM 61.24 0 0.00

FAMILY\_STATUS 61.24 0 0.00

max90days 61.85 1013 9.89

max60days 61.85 1448 14.14

max30days 61.85 1930 18.84

max21days 61.85 2315 22.60

max14days 61.85 2526 24.66

avg\_num\_delay 64.44 1569 15.32

if\_zalog 64.29 2427 23.69

num\_AccountActive180 64.29 2588 25.27

num\_AccountActive90 64.29 3115 30.41

num\_AccountActive60 64.29 3308 32.30

Active\_to\_All\_prc 64.29 492 4.80

numAccountActiveAll 64.29 479 4.68

numAccountClosed 64.29 404 3.94

sum\_of\_paym\_months 64.29 17 0.17

all\_credits 64.29 0 0.00

Active\_not\_cc 64.29 1210 11.81

own\_closed 64.29 2094 20.44

min\_MnthAfterLoan 64.29 165 1.61

max\_MnthAfterLoan 64.29 11 0.11

dlq\_exist 64.29 1585 15.47

thirty\_in\_a\_year 64.29 3125 30.51

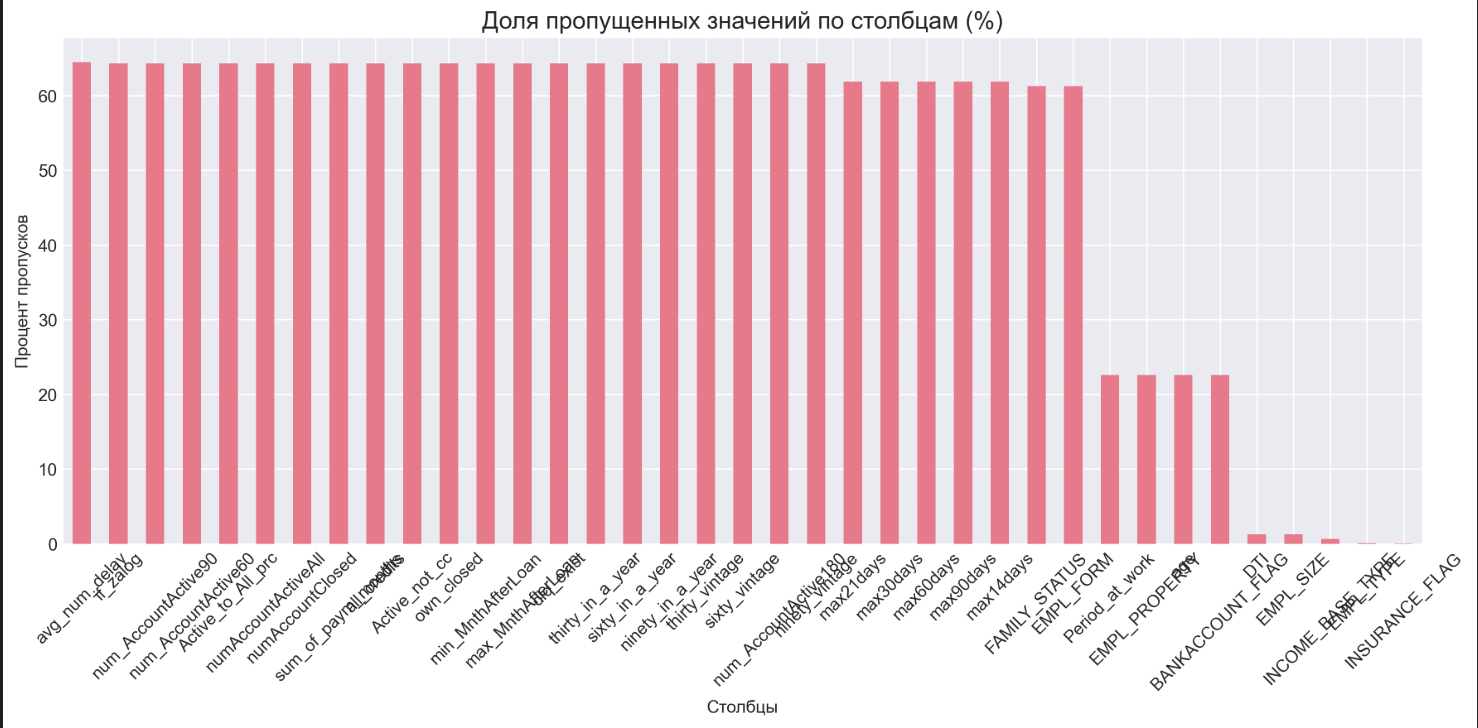
sixty\_in\_a\_year 64.29 3349 32.70

ninety\_in\_a\_year 64.29 3421 33.40

thirty\_vintage 64.29 3554 34.70

sixty\_vintage 64.29 3601 35.16

ninety\_vintage 64.29 3615 35.29

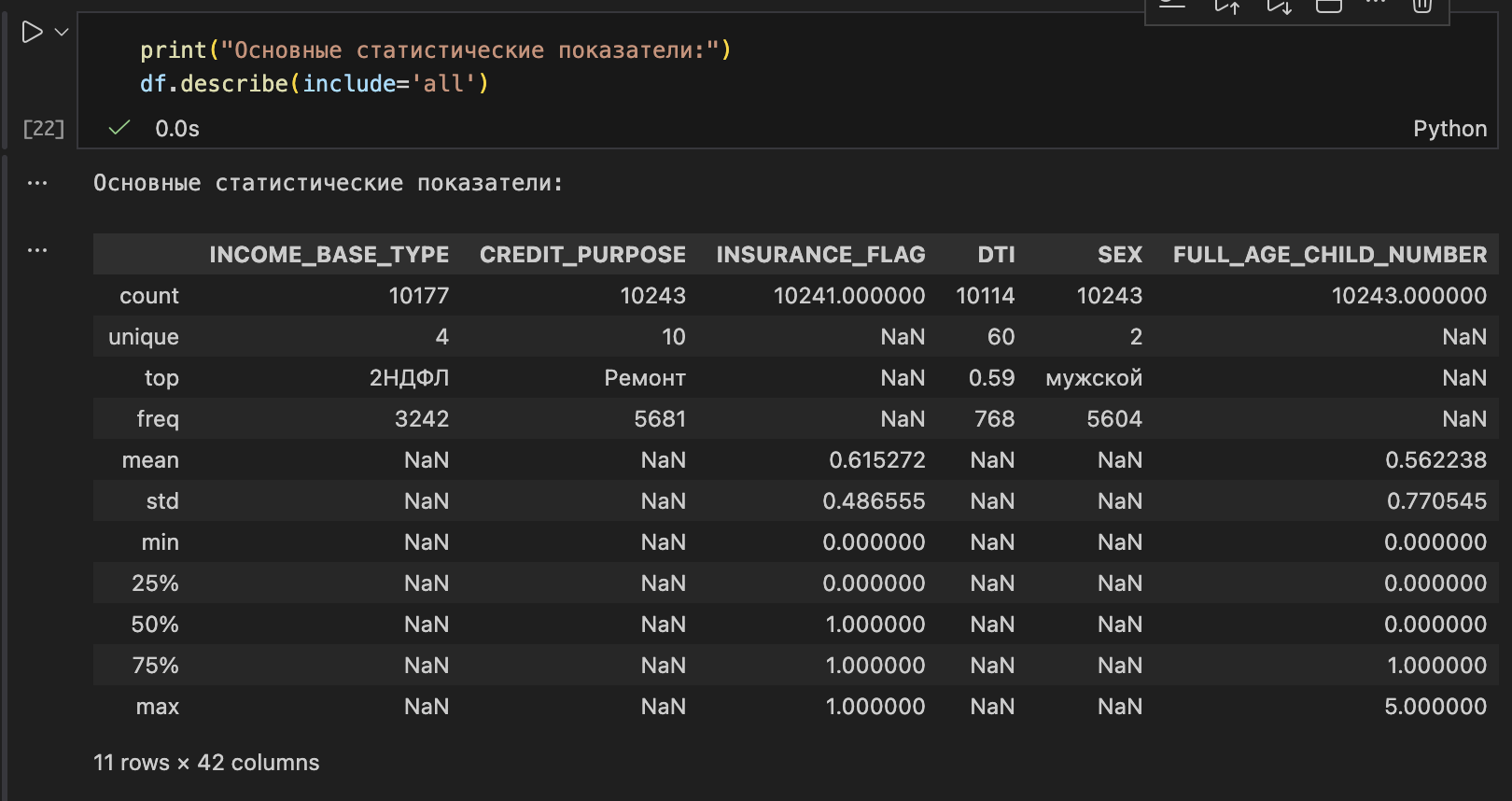


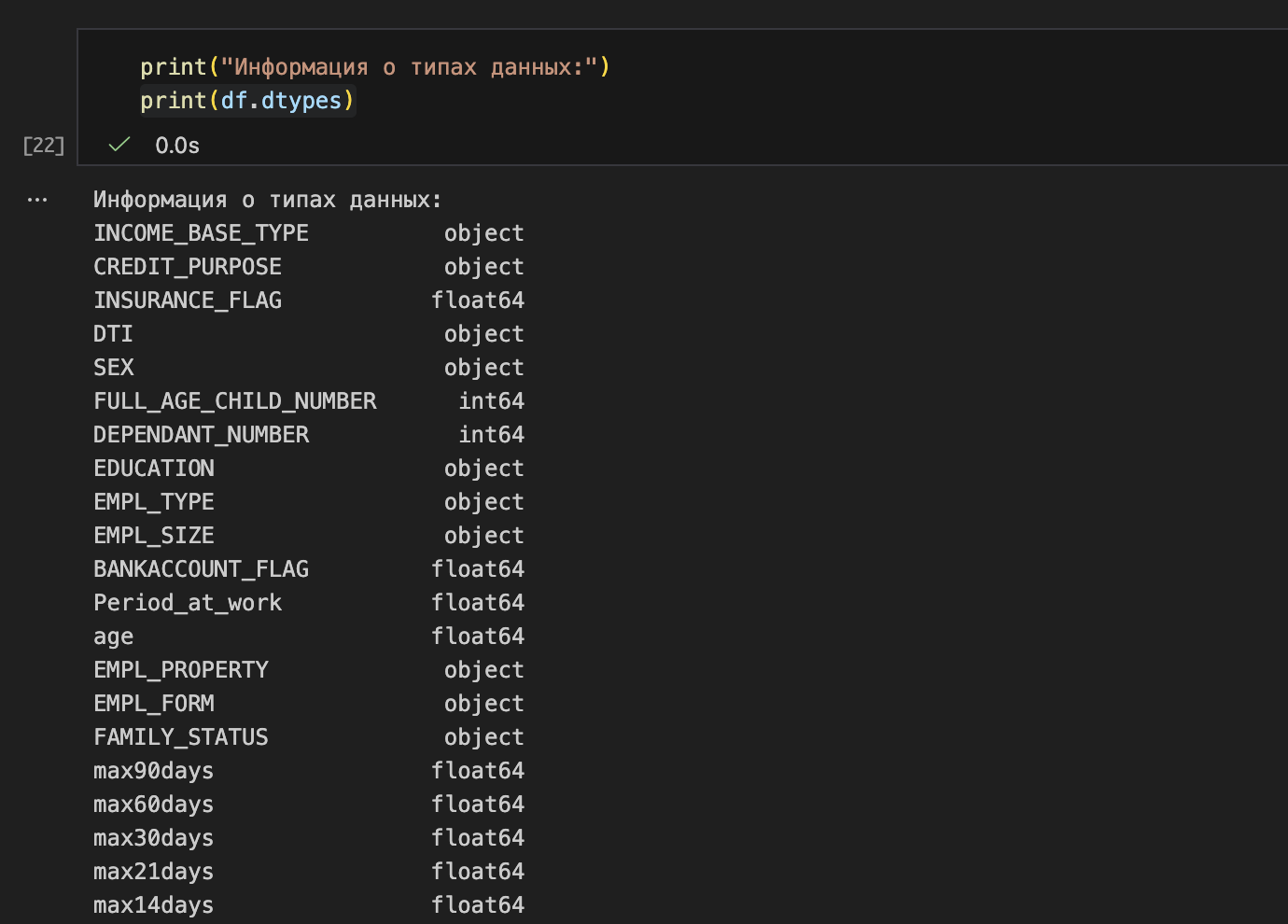
Из графика видим что три четверти переменных имеют долю пропущенных более 60%.

Доля нулевых для половины переменных составляет около 20-30%, для другой половины близка к нулю. Исключением является DEPENDANT\_NUMBER для которой доля нулевых составляет 99.7%.

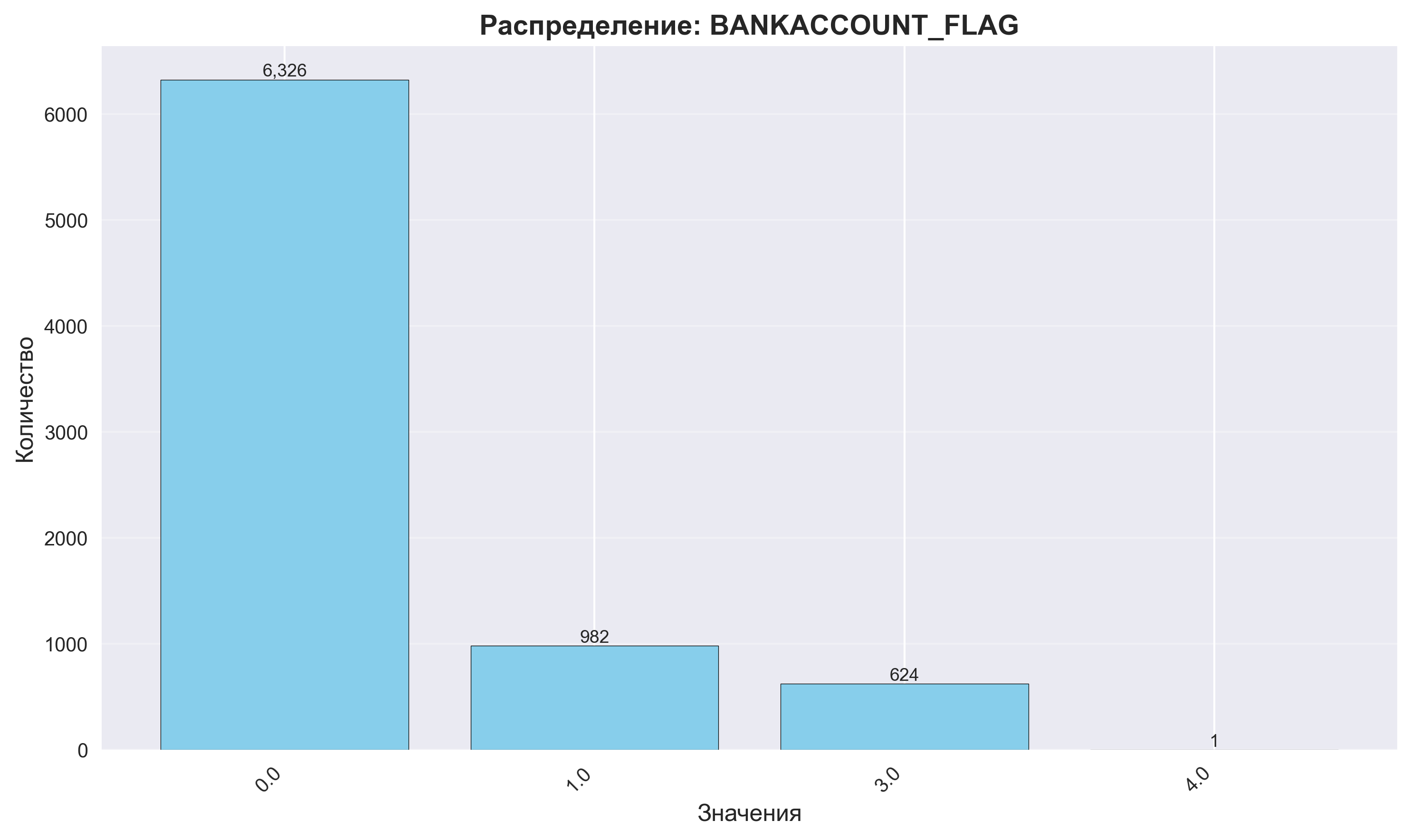
* Среднее значение, медиана, стандартное отклонение, минимум, максимум, тип данных по каждому показателю в предоставленных данных;

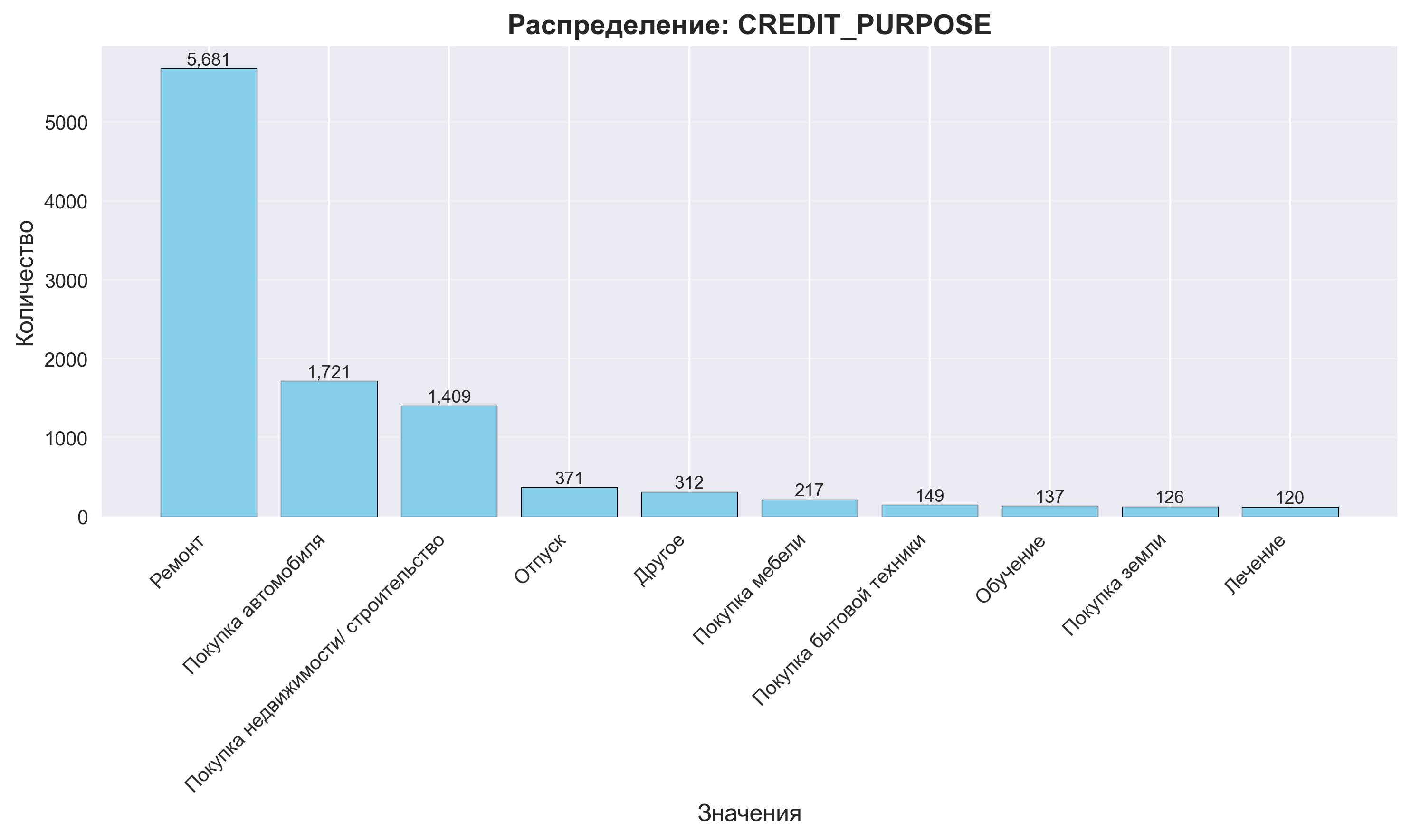
(Полные данные в файле task1.ipynb)

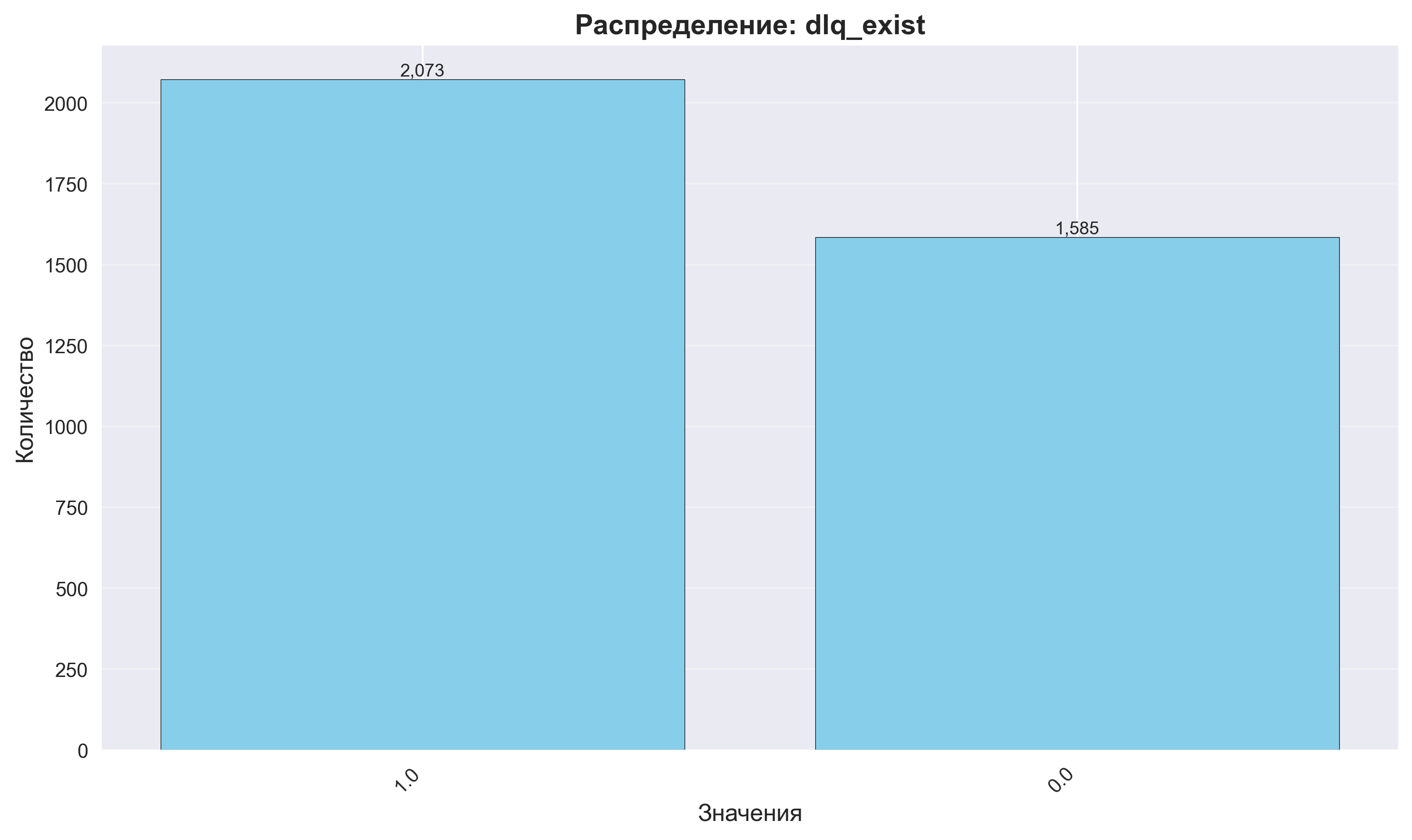


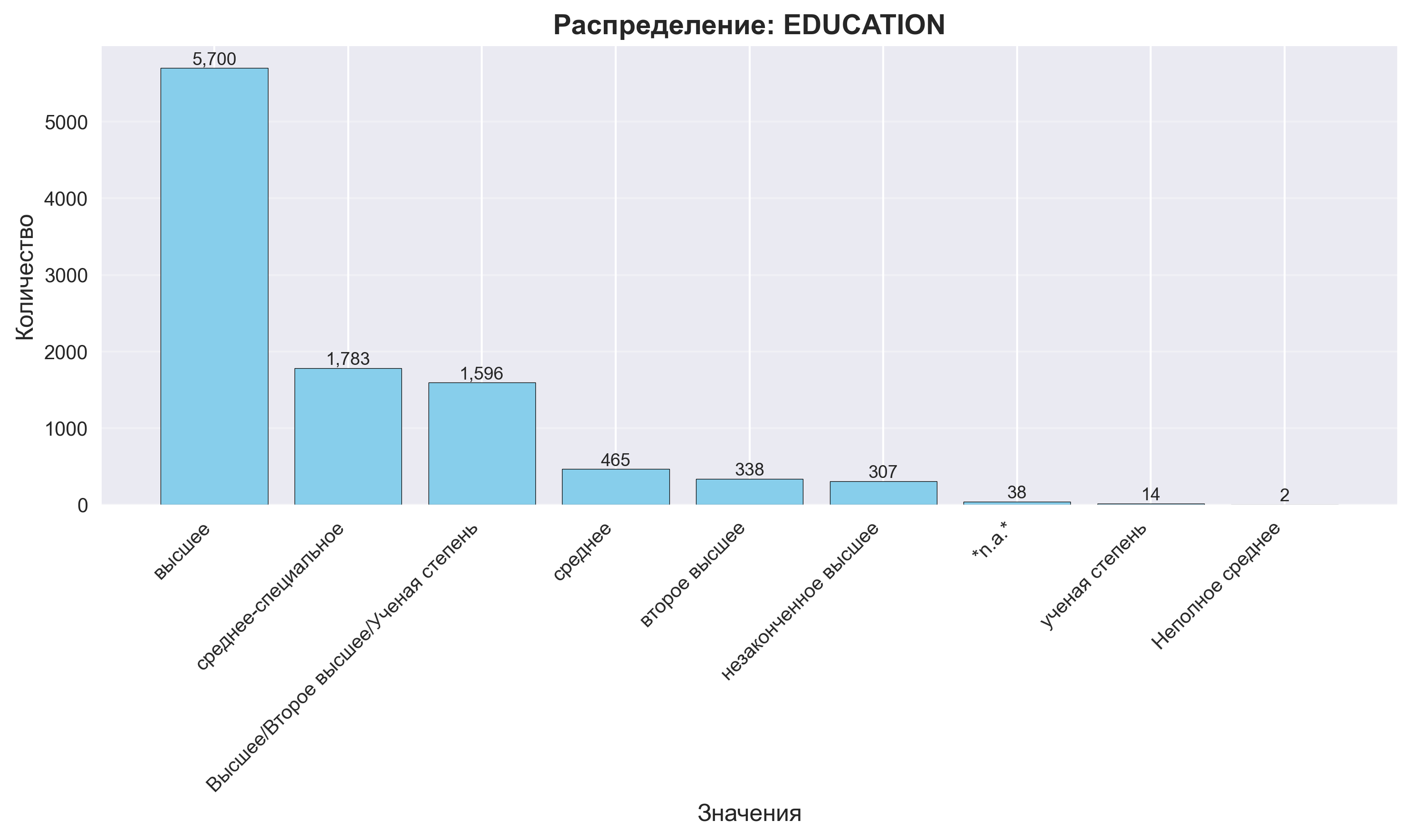


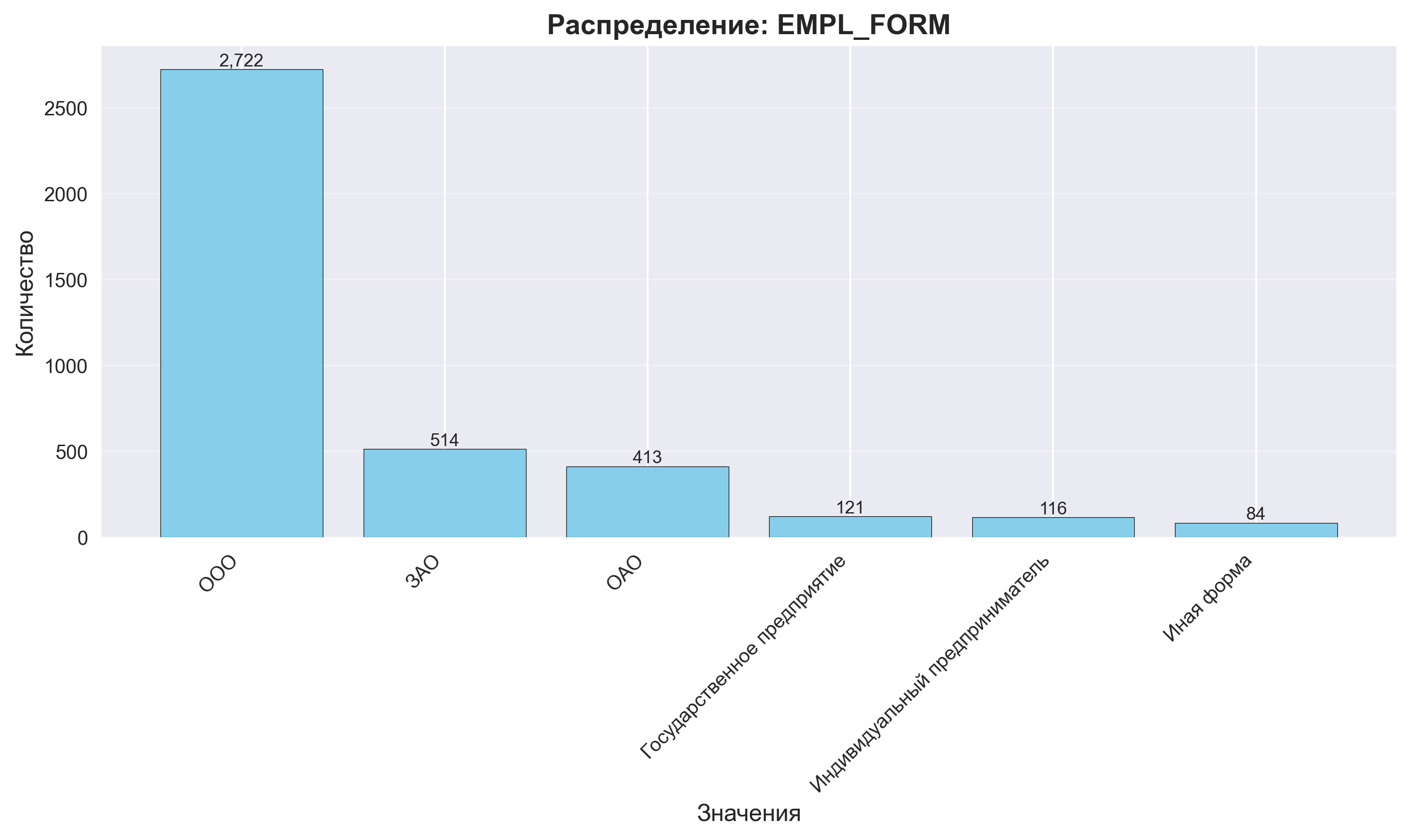
* Исследуем распределение данных по полу, возрасту и другим категориальным показателям;

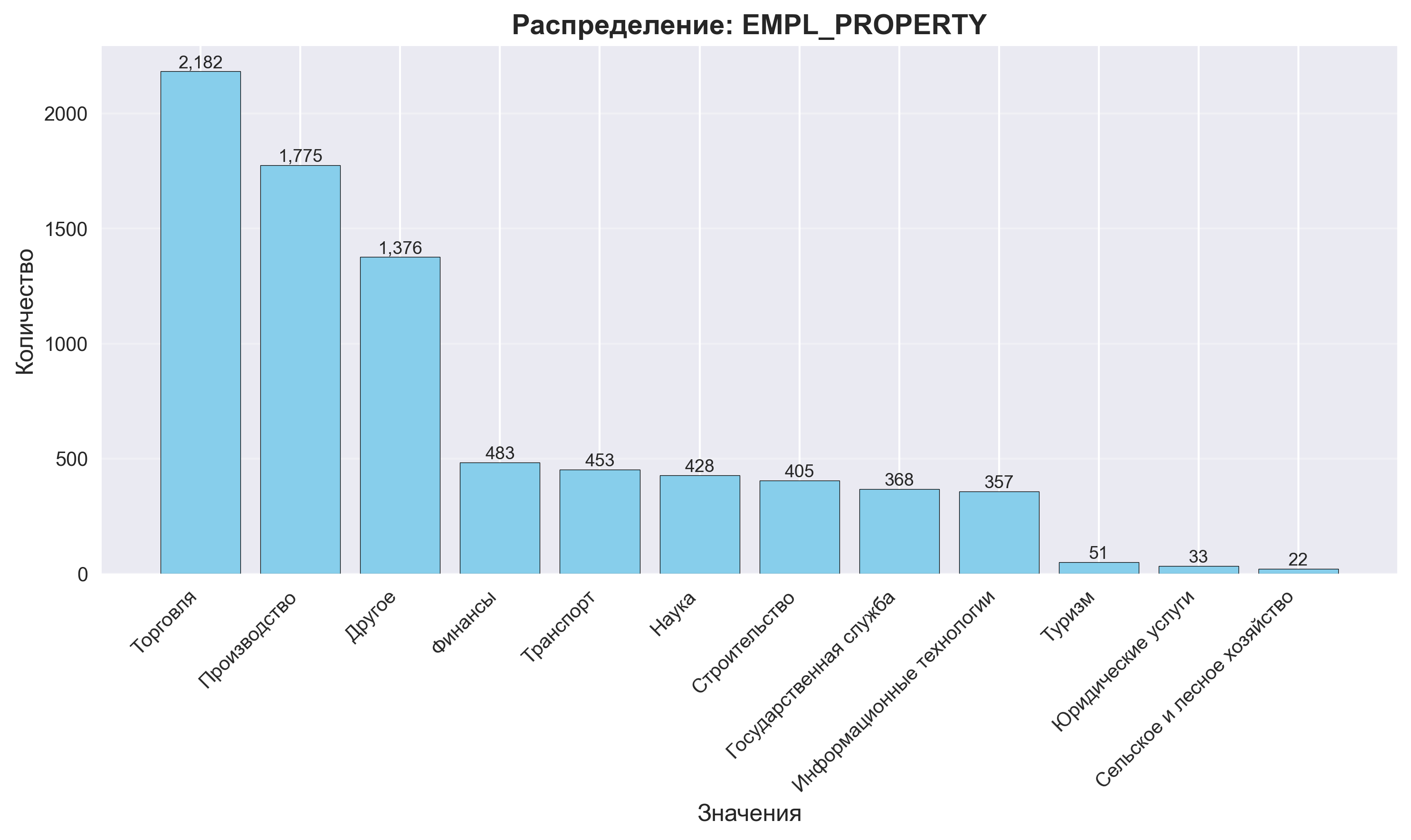


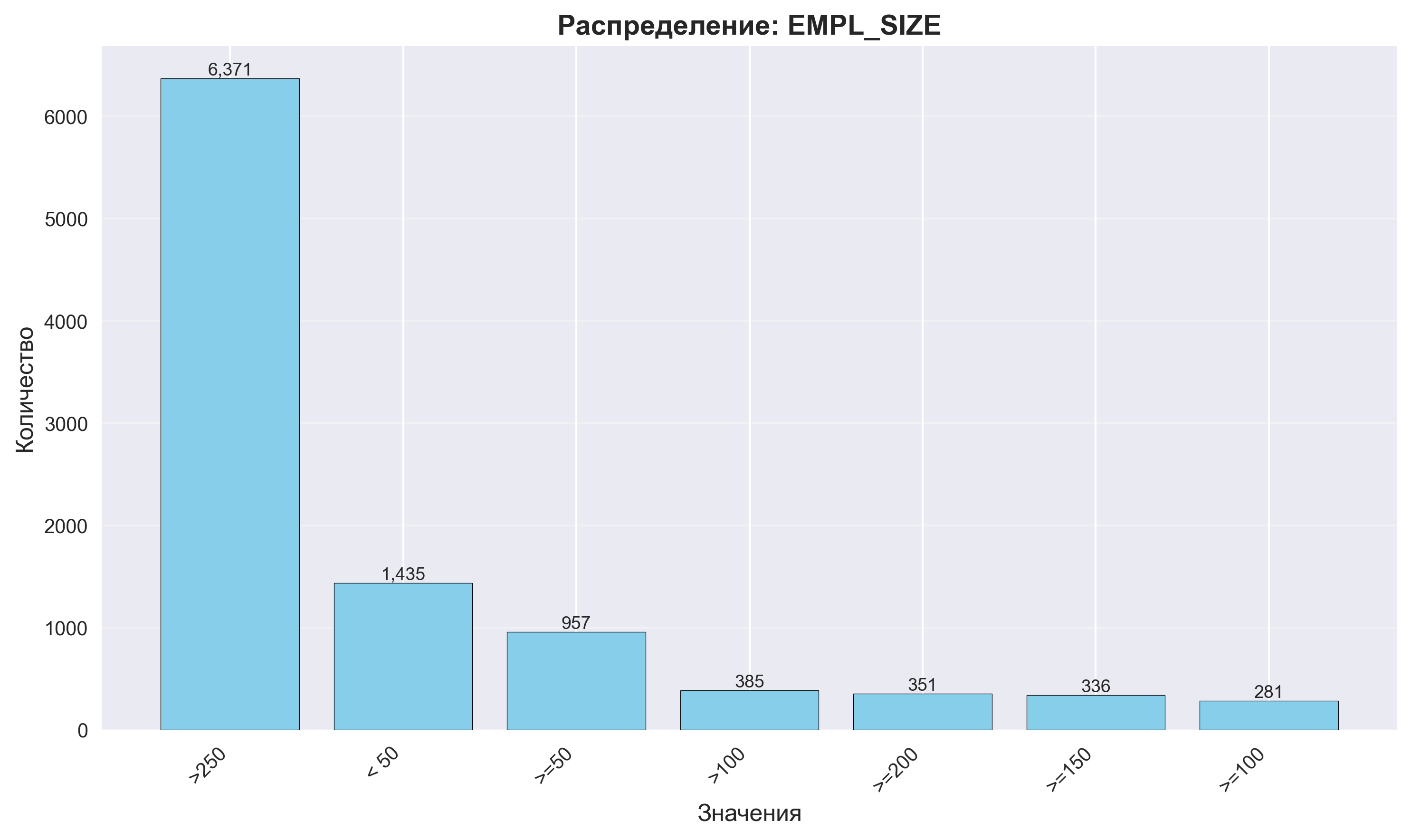


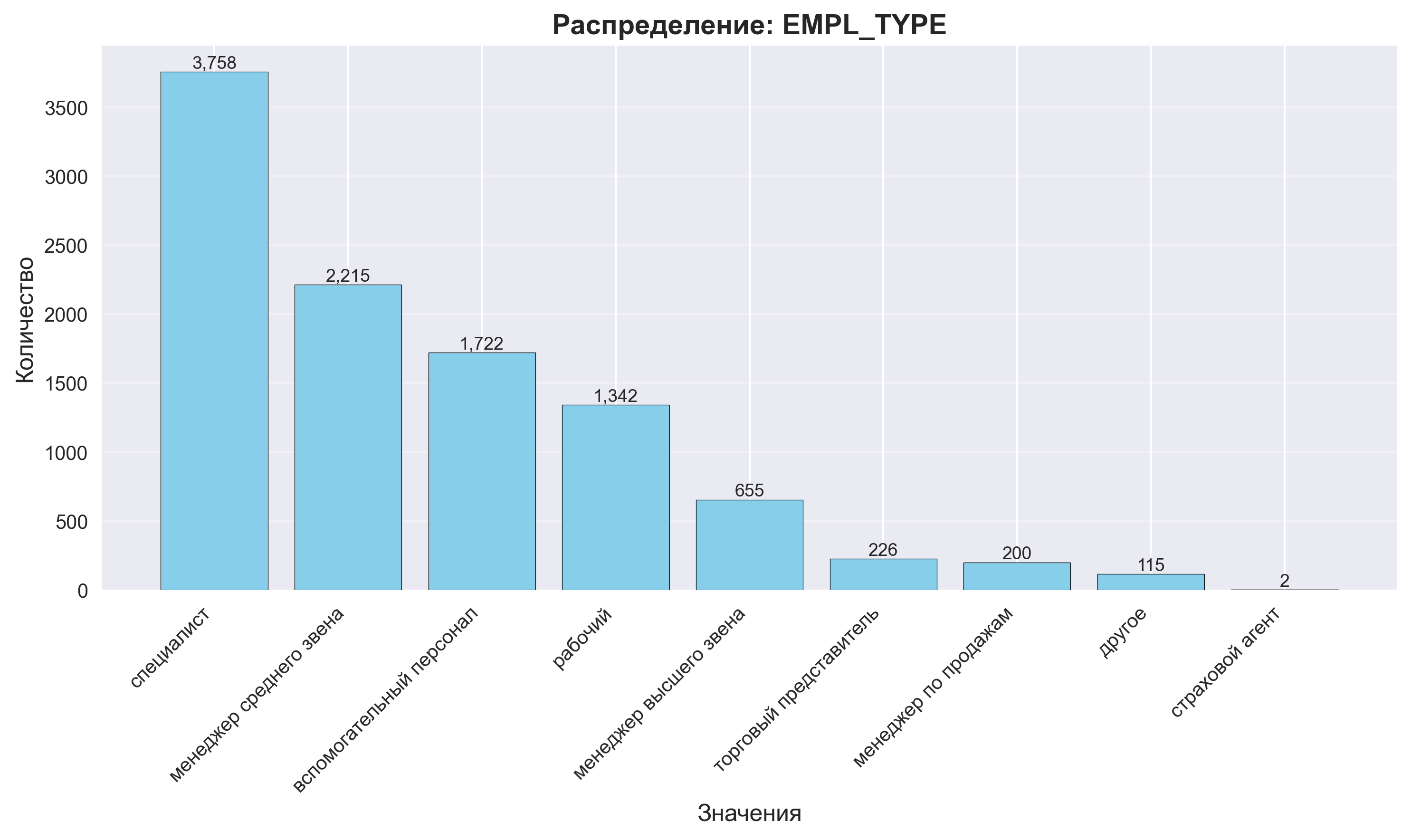


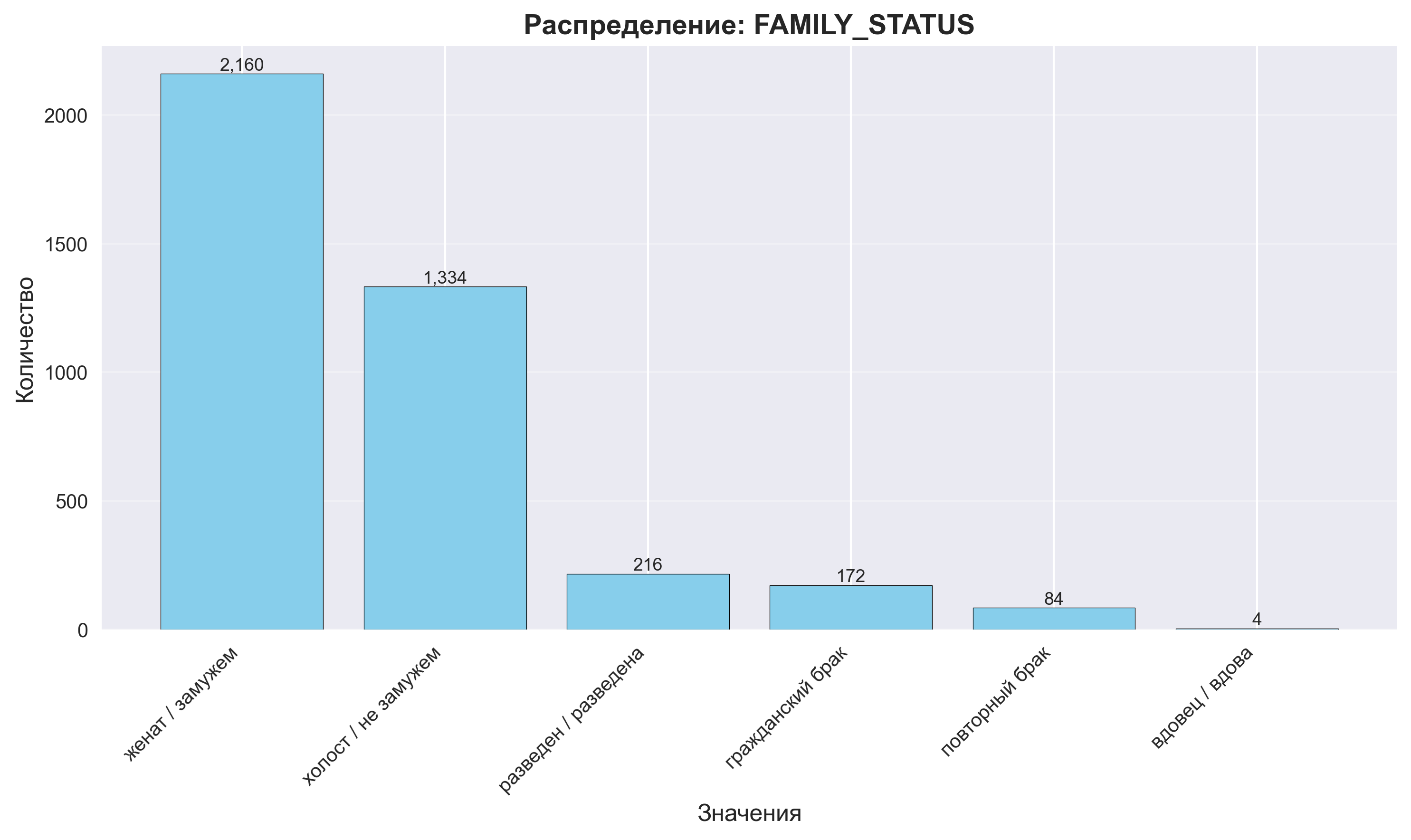


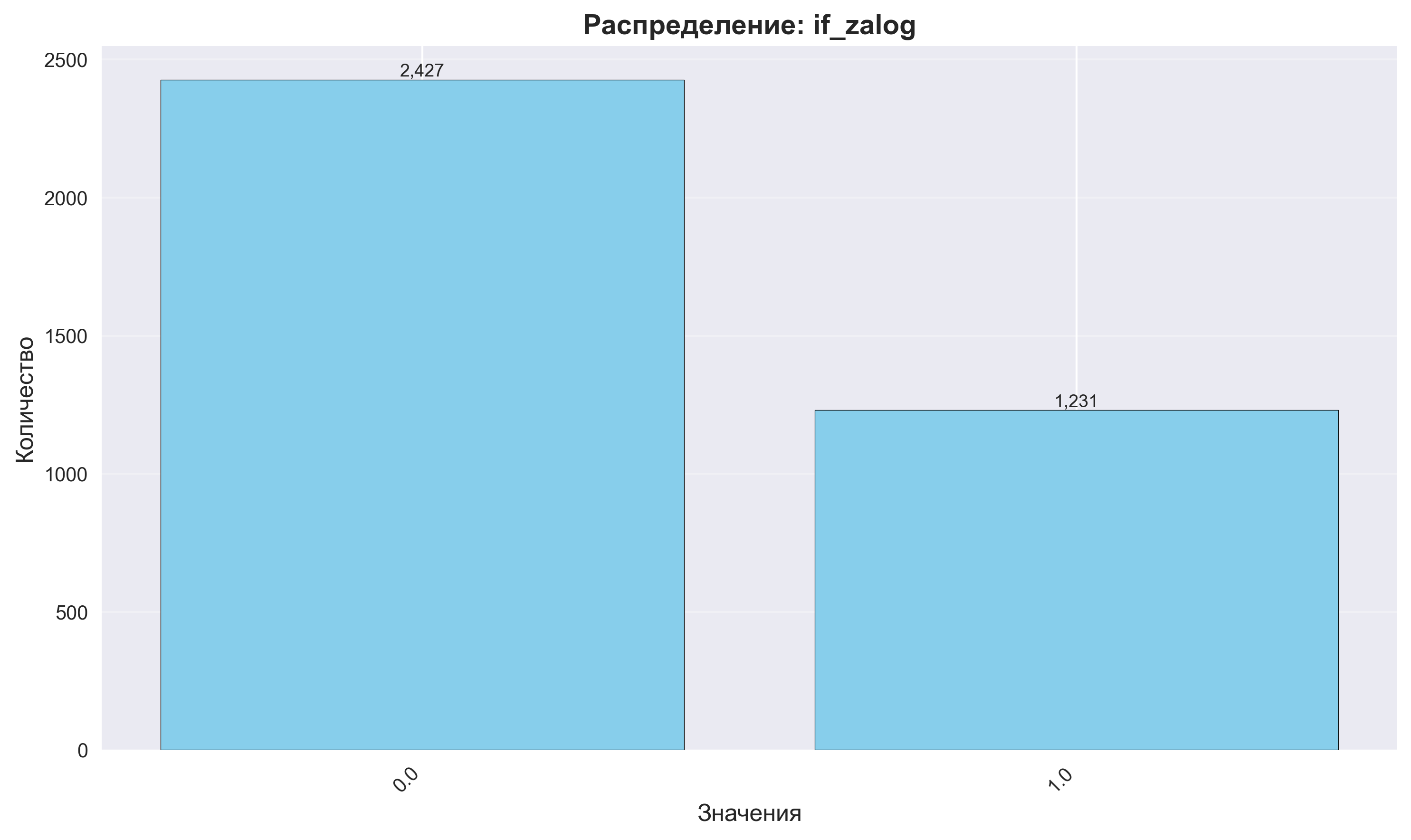


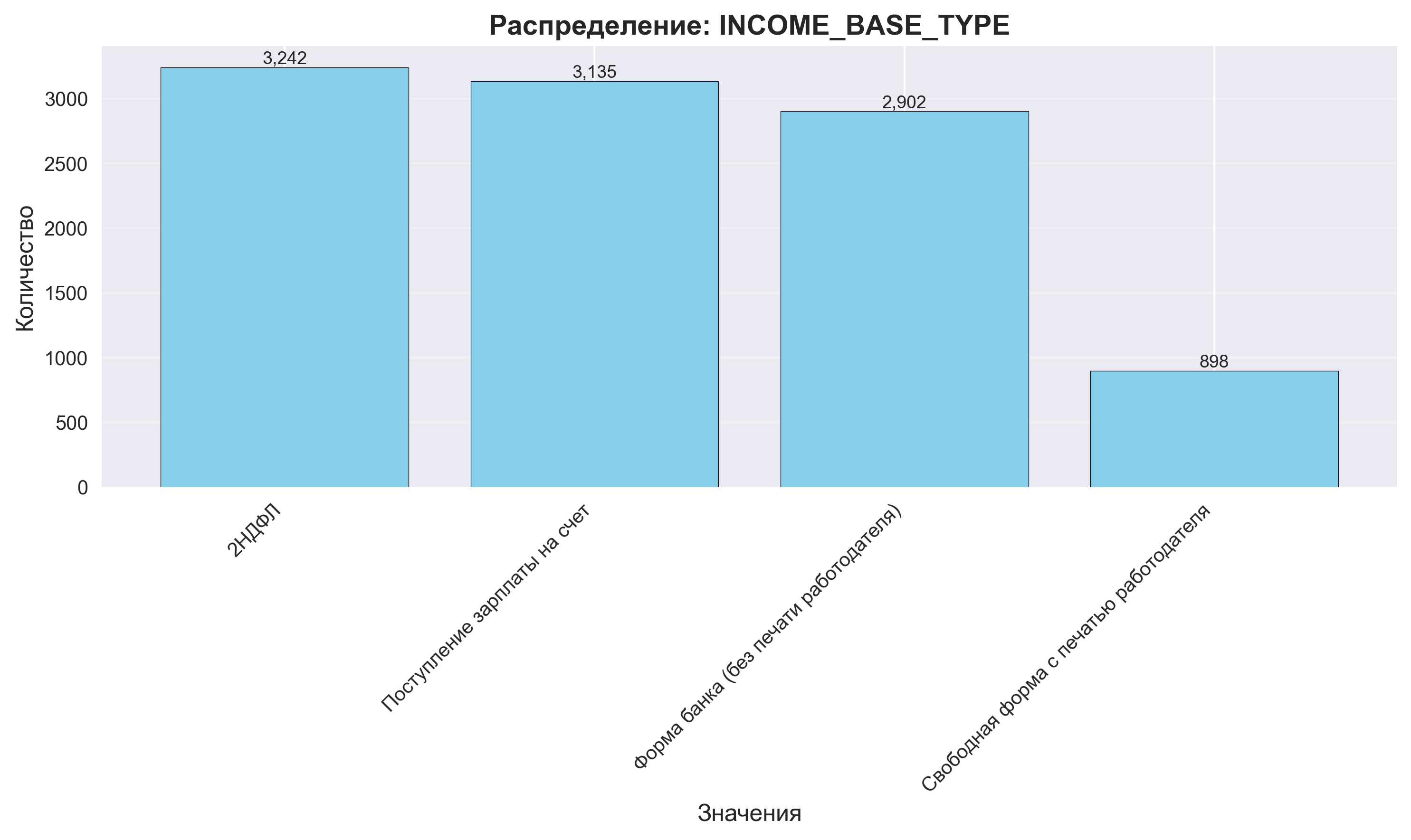


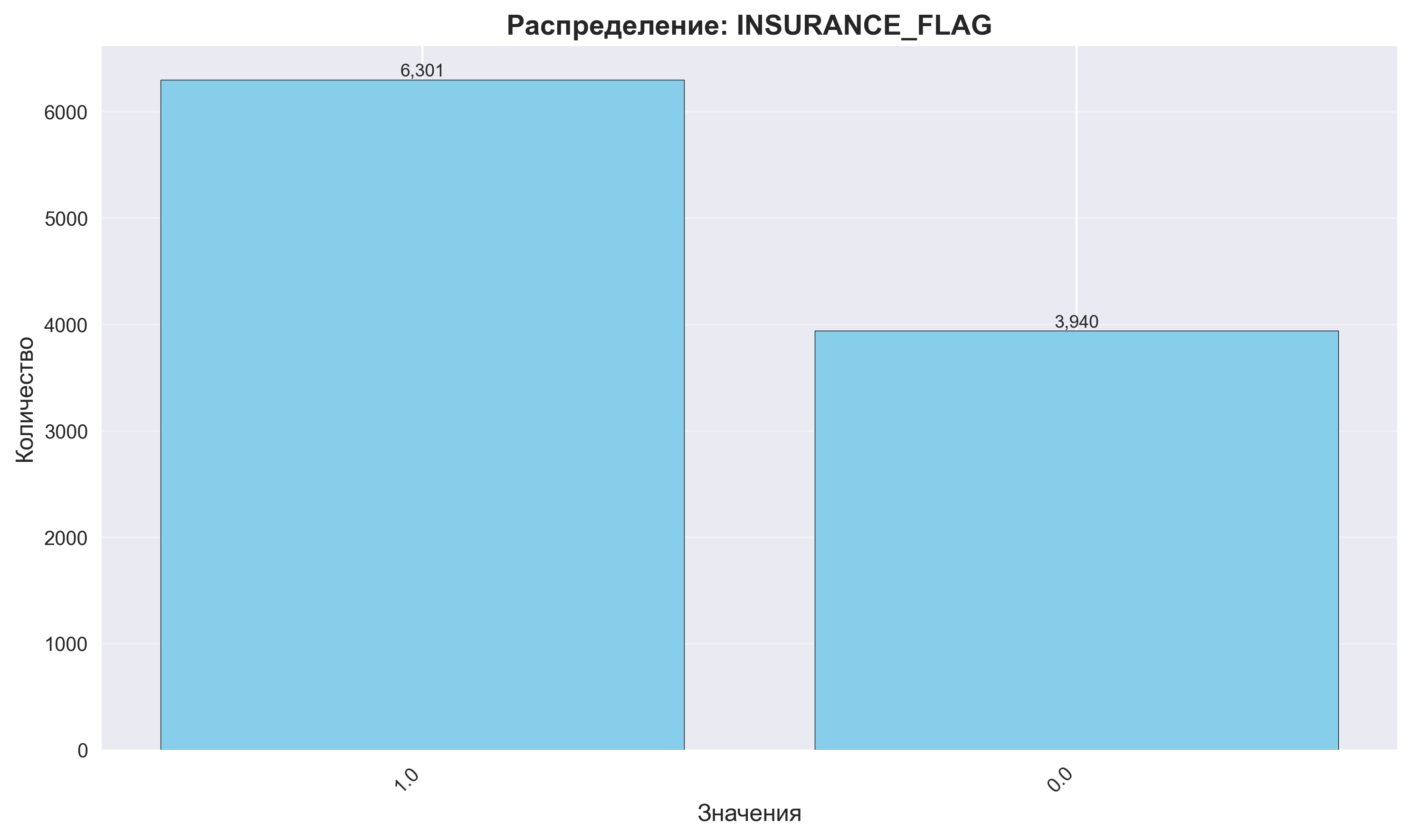


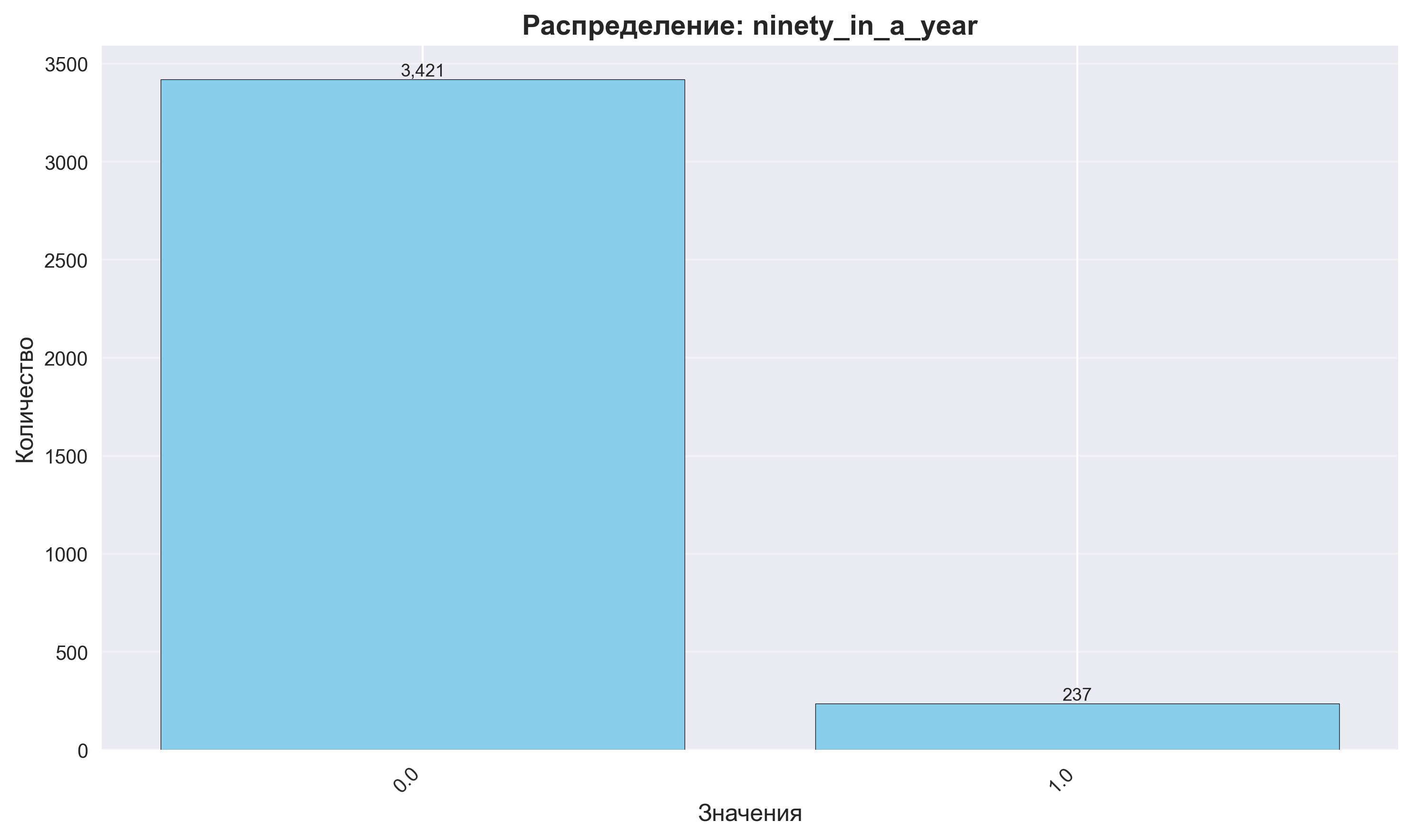




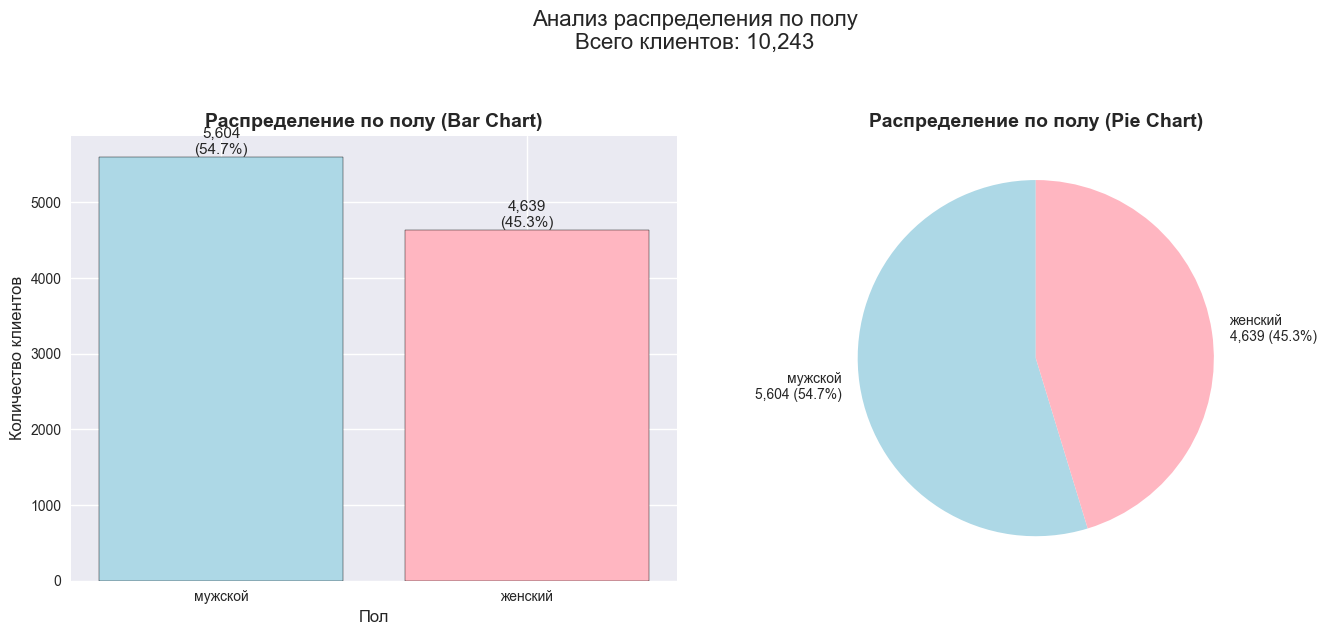


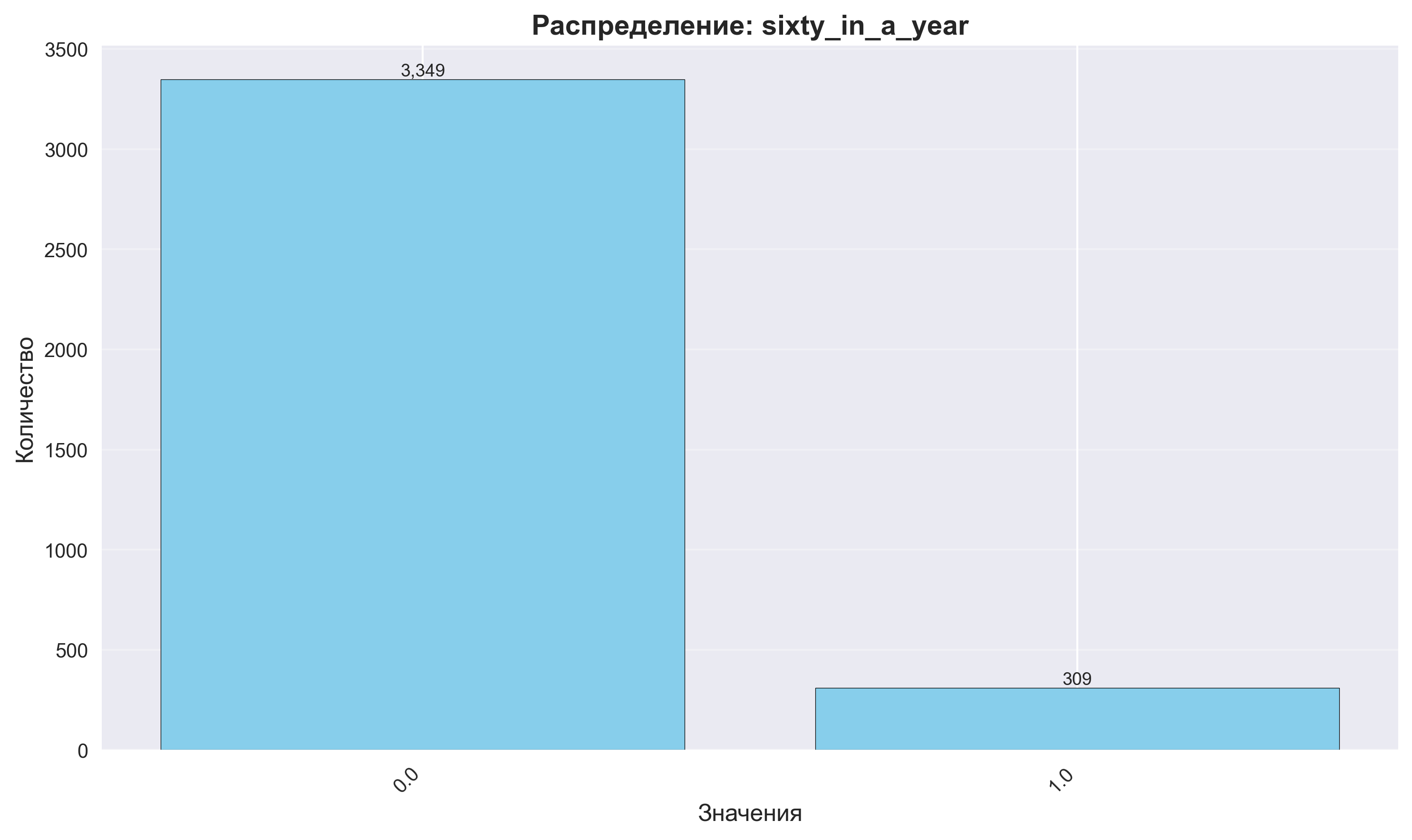


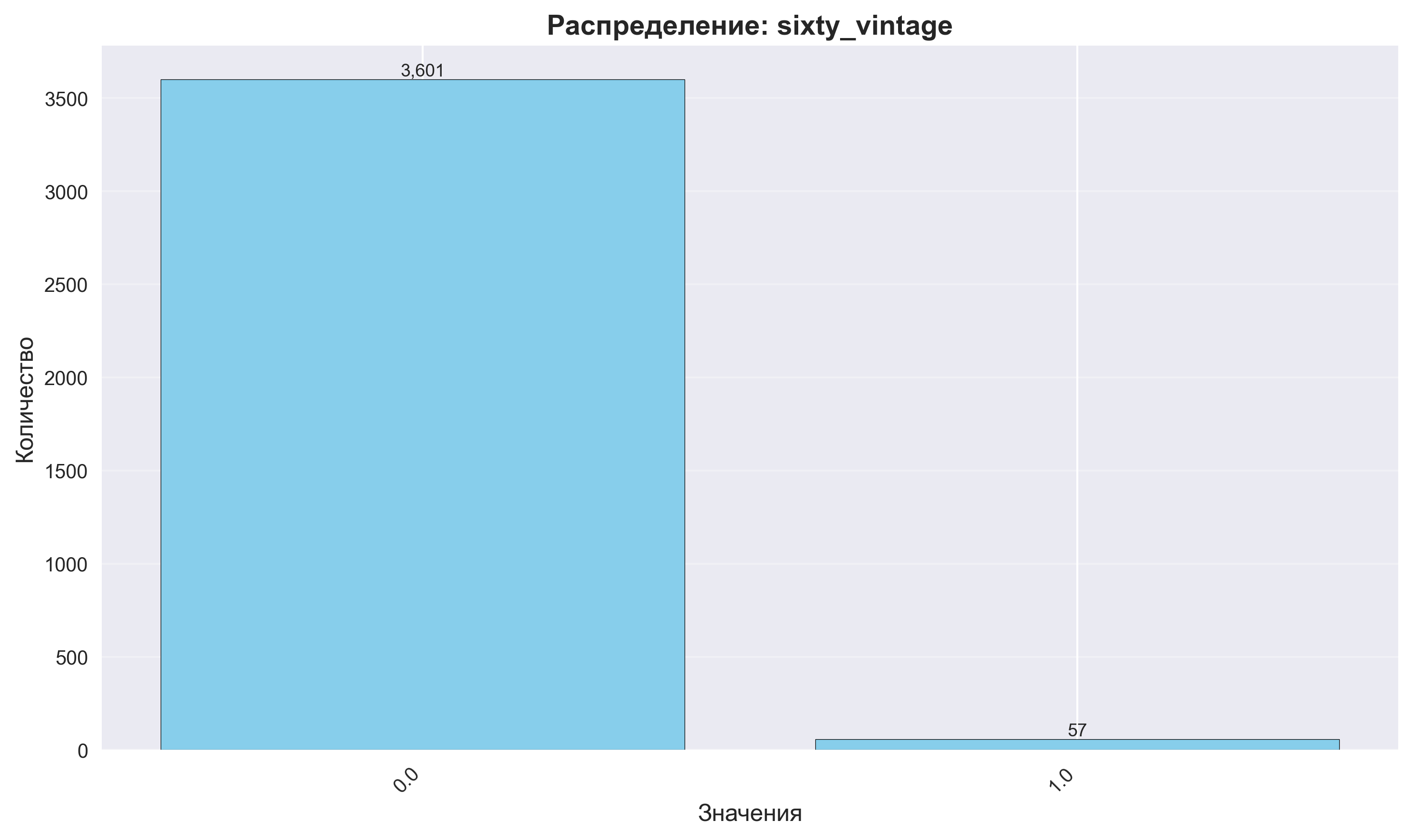


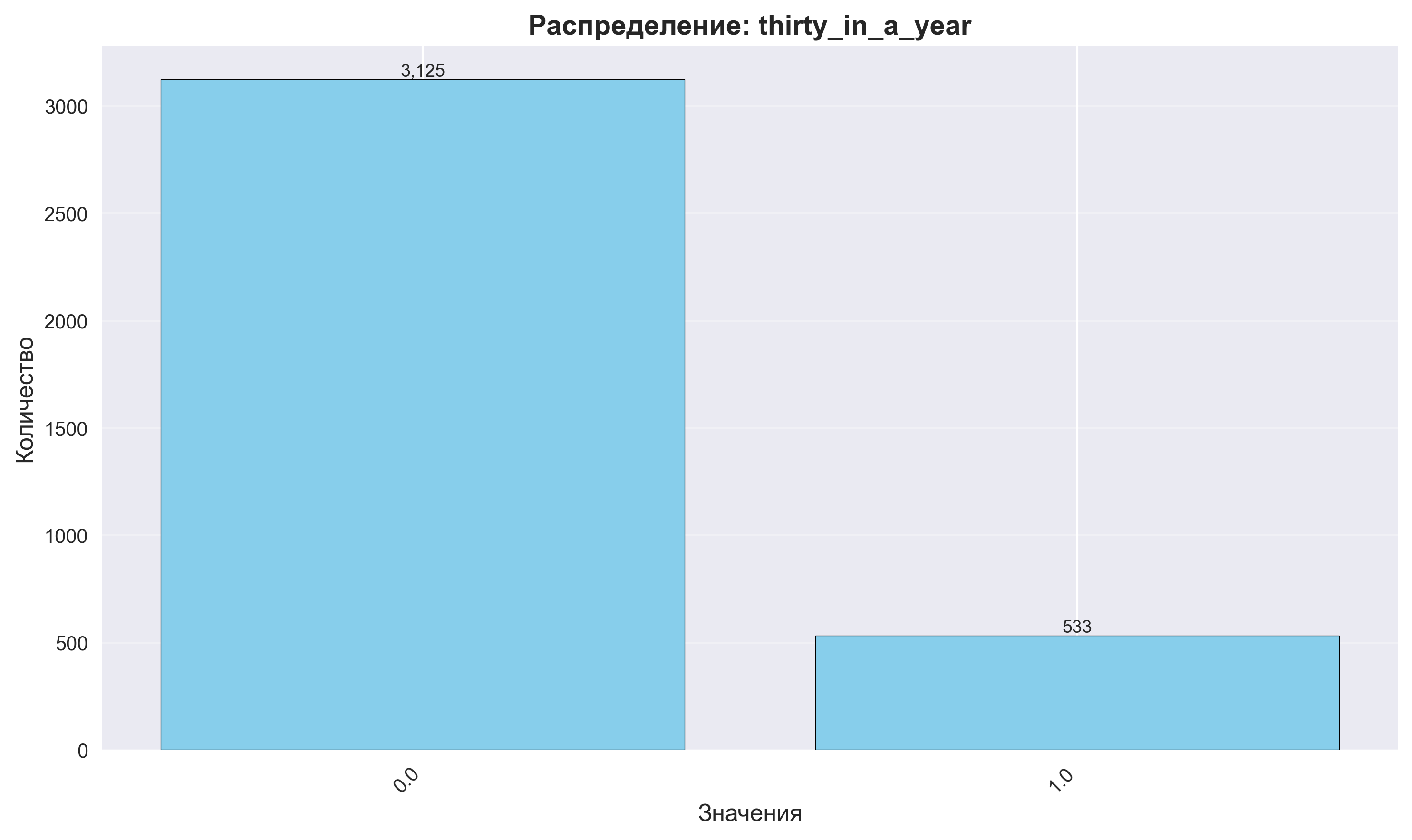


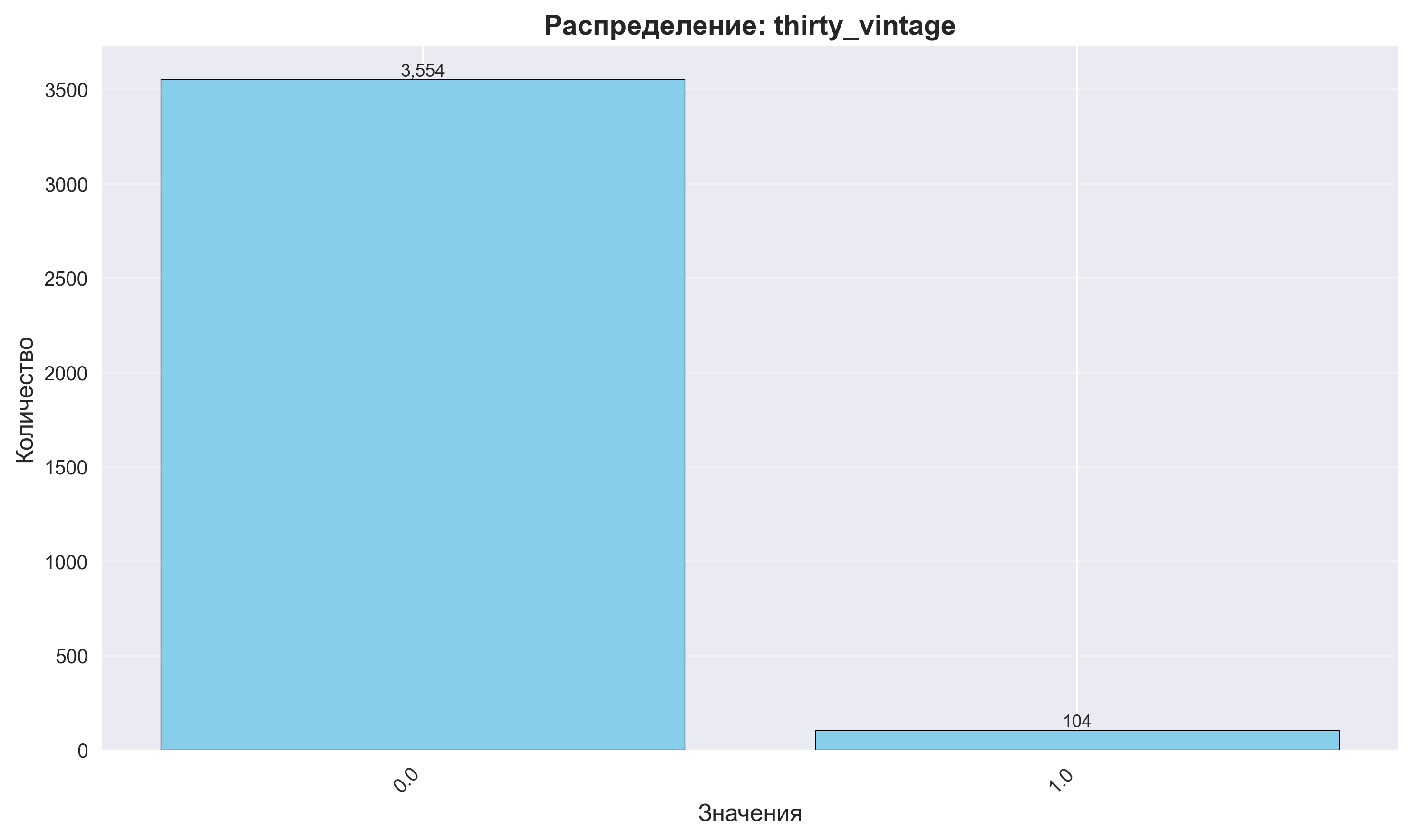












* Выполним проверку на наличие некорректных знаков с помощью следующей регулярки r'[^a-zA-Zа-яА-Я0-9\s/()<>=.,!?-]'

Итого строки с некорректными символами:

DTI: 1 строк с некорректными символами

EDUCATION: 38 строк с некорректными символами

в EDUCATION это \*n.a.\* - аббревиатура нет данных, заменим её на null значение.

в DTI это единственное значение #ПОЛЕ! заменим его на среднее для DTI - 0.39

## Часть 1.2 Исследование переменных

**Категориальные переменные:**

- **INCOME\_BASE\_TYPE** (Подтверждение дохода) принимает 5 различных значений, включая Null в случае пропуска.

- **EDUCATION** (Образование) принимает 9 различных значений уровня образования. Однако значение \*n.a\* нужно удалить вручную.

- **EMPL\_TYPE** (Должность) – принимает 10 различных значений, включая значение Null в случае пропуска.

- **EMPL\_SIZE** (Зарплата) – принимает 8 различных значений, включая значение NULL, в случае пропуска, однако пару значений '>100', '>=100' мы можем объединить в одну '>=100'

- **BANKACCOUNT\_FLAG** (Кол-во аккаунтов у клиента) принимает 5 различных значений (от 0 до 4 и пропуск). Но значения 2, 3 и 4 означают одно и то же, так что мы заменим их на 2.

- **CREDIT\_PURPOSE** (Цель получения кредита) принимает 10 различных значений, без пропусков.

- **SEX** (Пол) может принимать 2 значений (мужской/женский), пропуски отсутствуют

- **EMPL\_PROPERTY** (Сфера бизнеса работодателя) – 12 уникальных значений + пропуски (2310 шт)

- **EMPL\_FORM** (Организационно - правовая форма) – 6 уникальных значений + пропуски (6273 шт)

- **FAMILY\_STATUS** (Семейный статус) – 6 уникальных значений + пропуски (6273 шт)

- **INSURANCE\_FLAG** (Страхование заемщика при получении кредита) - true/false, 2 пропуска

- **if\_zalog** (Наличие залога (квартира,машина)) – 6585 пропусков

- **dlq\_exist** (Наличие просрочки на данный момент) – 6585 пропусков

- **thirty\_in\_a\_year** (Просрочка больше 30 дней за последний год) – 6585 пропусков

- **sixty\_in\_a\_year** (Просрочка больше 60 дней за последний год) – 6585 пропусков

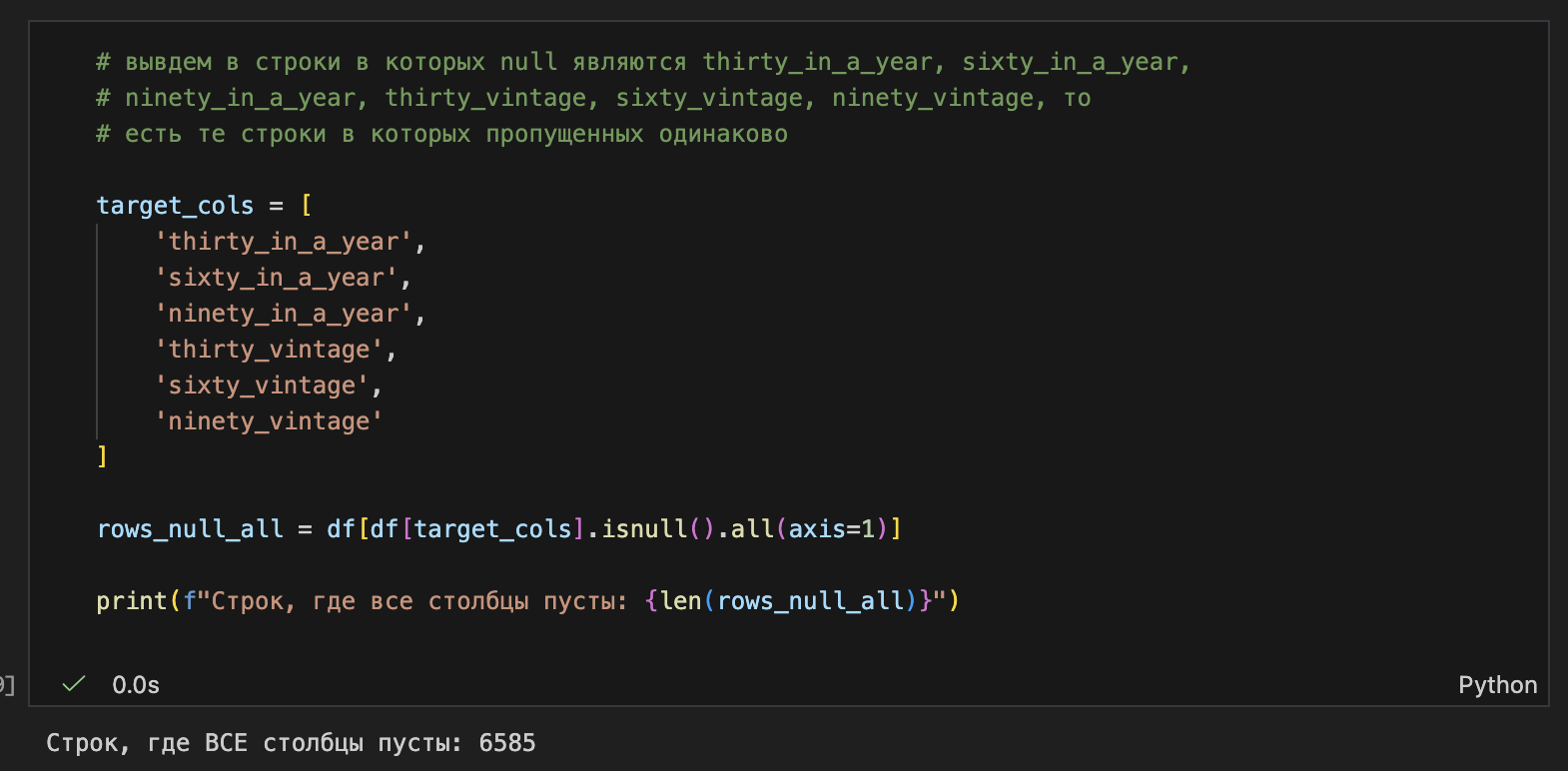
- **ninety\_in\_a\_year** (Просрочка больше 90 дней за последний год) – 6585 пропусков

- **thirty\_vintage** (Просрочка больше 30 дней, когда-либо) – 6585 пропусков

- **sixty\_vintage** (Просрочка больше 60 дней, когда-либо) – 6585 пропусков

- **ninety\_vintage** (Просрочка больше 90 дней, когда-либо) – 6585 пропусков

В случае с категориальными переменными эффективнее удалить строки с пропущенными значениями. Предварительно проверим не совпдают ли эти строки в которых по 6585 пропусков для последних 8 переменных.



Отлично все строки совпали, тогда удалим эти строки.

Посмотрим на null строки по новому датафрейму  
 Тип данных Уникальных значений Пустых значений

INCOME\_BASE\_TYPE object 4 0

CREDIT\_PURPOSE object 10 0

INSURANCE\_FLAG float64 2 1

SEX object 2 0

EDUCATION object 8 16

EMPL\_TYPE object 8 0

EMPL\_SIZE object 2 0

BANKACCOUNT\_FLAG float64 3 0

EMPL\_PROPERTY object 5 0

EMPL\_FORM object 6 0

FAMILY\_STATUS object 6 0

if\_zalog float64 2 0

dlq\_exist float64 2 0

thirty\_in\_a\_year float64 2 0

sixty\_in\_a\_year float64 2 0

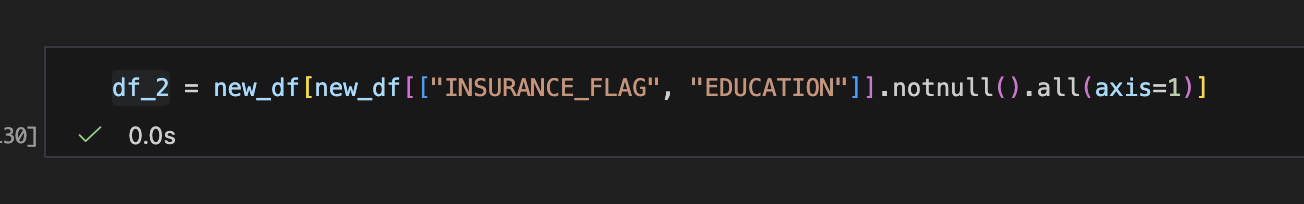
ninety\_in\_a\_year float64 2 0

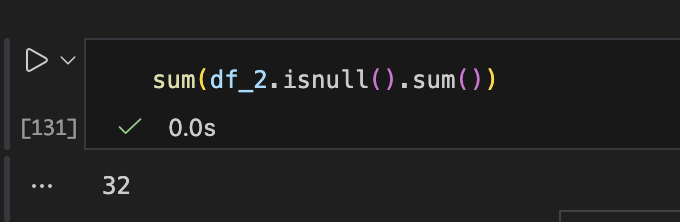
thirty\_vintage float64 2 0

sixty\_vintage float64 2 0

ninety\_vintage float64 2 0

Удалим также INSURANCE\_FLAG и EDUCATION с пустым значением.

  
Теперь null значения остались только у количественных переменных. Для них мы применим другой метод - заменим null на медианное значение.



Всего таких null значений 32, а это менее 1% количества оставшихся строк, так что такая замена не сильно испортит данные, при этом эти 32 значения не находятся в одном столбце.

Закодируем категориальные переменные в целочисленные. Предворительно заменим категорию "Высшее/Второе высшее/Ученая степень" в переменной EDUCATION на "Высшее". Будем считать замену честной т.к. среди категорий ["Высшее", "Второе высшее", "Ученая степень"] "Высшее" встречается чаще всего и при этом является минимальной из них.

**Количественные переменные:**   
- **DTI** (debt-to-income ratio — отношение долга к доходам)

- **sum\_of\_paym\_months** (Сумма платежей за последний месяц (тыс.))

- **FULL\_AGE\_CHILD\_NUMBER** (Кол-во лет ребенку)

- **all\_credits** (Кол-во кредитов)

- **Active\_not\_cc** (Активные кредитные счета, но не кредитная карта)

- **DEPENDANT\_NUMBER** (Кол-во иждивенцев)

- **own\_closed** (Кол-во закрытых кредитов)

- **min\_MnthAfterLoan** (Минимальное кол-во месяцев, которое

прошло с момета взятия прошлого кредита

- **Period\_at\_work** (Время работы (кол-во дней))

- **age** (Возраст)

- **max\_MnthAfterLoan** (Кол-во месяцев прошеднее с момента выдачи первого кредита)

- **avg\_num\_delay** (Среднее кол-во задержки оплаты)

- **num\_AccountActive180** (Кол-во активных счетов счетов за последние

180 дней)

- **num\_AccountActive90** (Кол-во активных счетов счетов за последние

90 дней)

- **num\_AccountActive60** (Кол-во активных счетов счетов за последние

60 дней)

- **Active\_to\_All\_prc** (Отношение активных счетов ко всем счетам)

- **numAccountActiveAll** (Кол-во открытых счетов)

- **numAccountClosed** (Кол-во закрытых счетов)

- **max90days** (Кол-во запросов в бюро кредитных историй за

последние 90 дней)

- **max60days** (Кол-во запросов в бюро кредитных историй за

последние 60 дней)

- **max30days** (Кол-во запросов в бюро кредитных историй за

последние 30 дней)

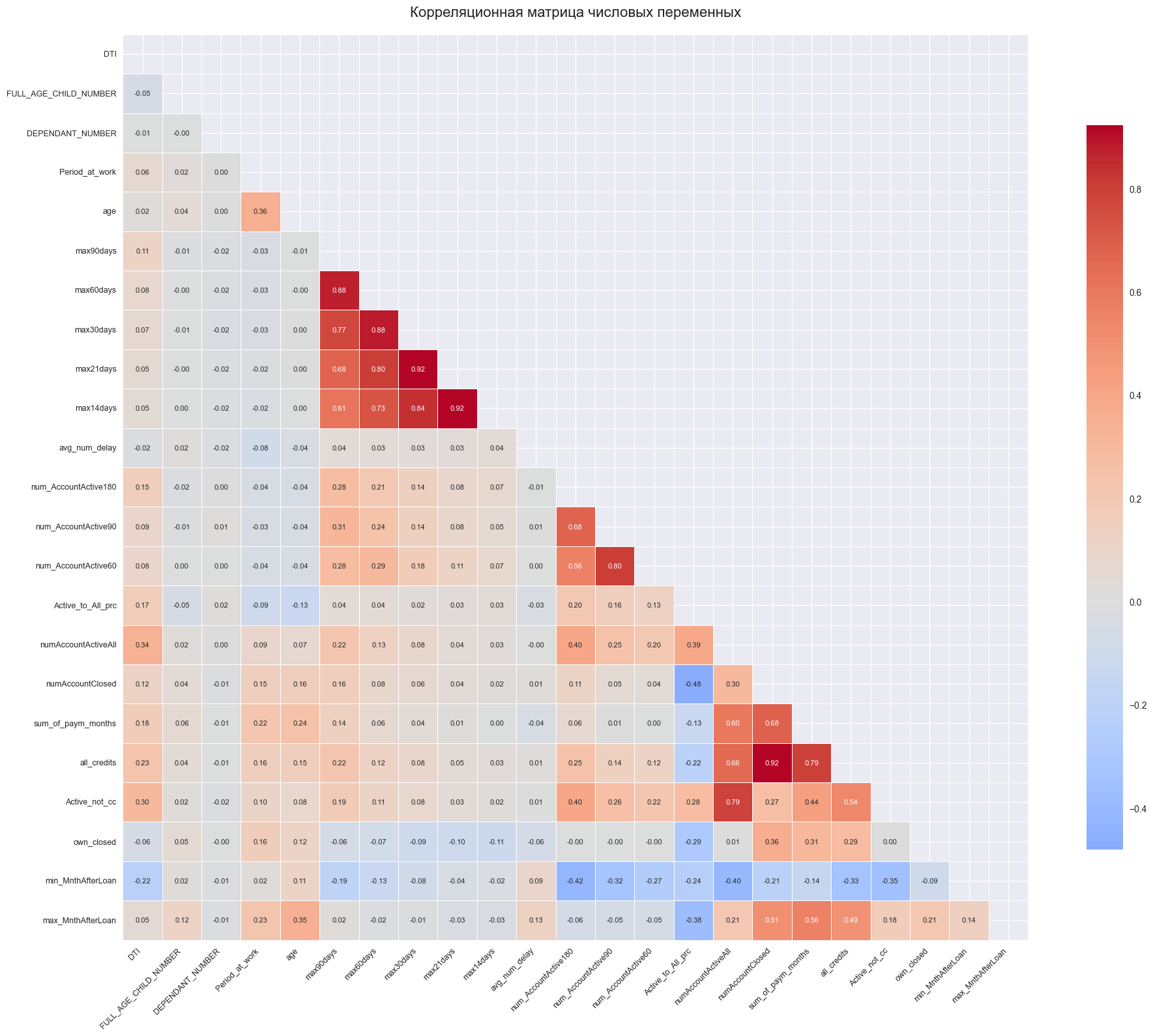
- **max21days** (Кол-во запросов в бюро кредитных историй за

последние 21 дней)

- **max14days** (Кол-во запросов в бюро кредитных историй за

последние 14 дней)

Построим матрицу корреляции для количественных переменных



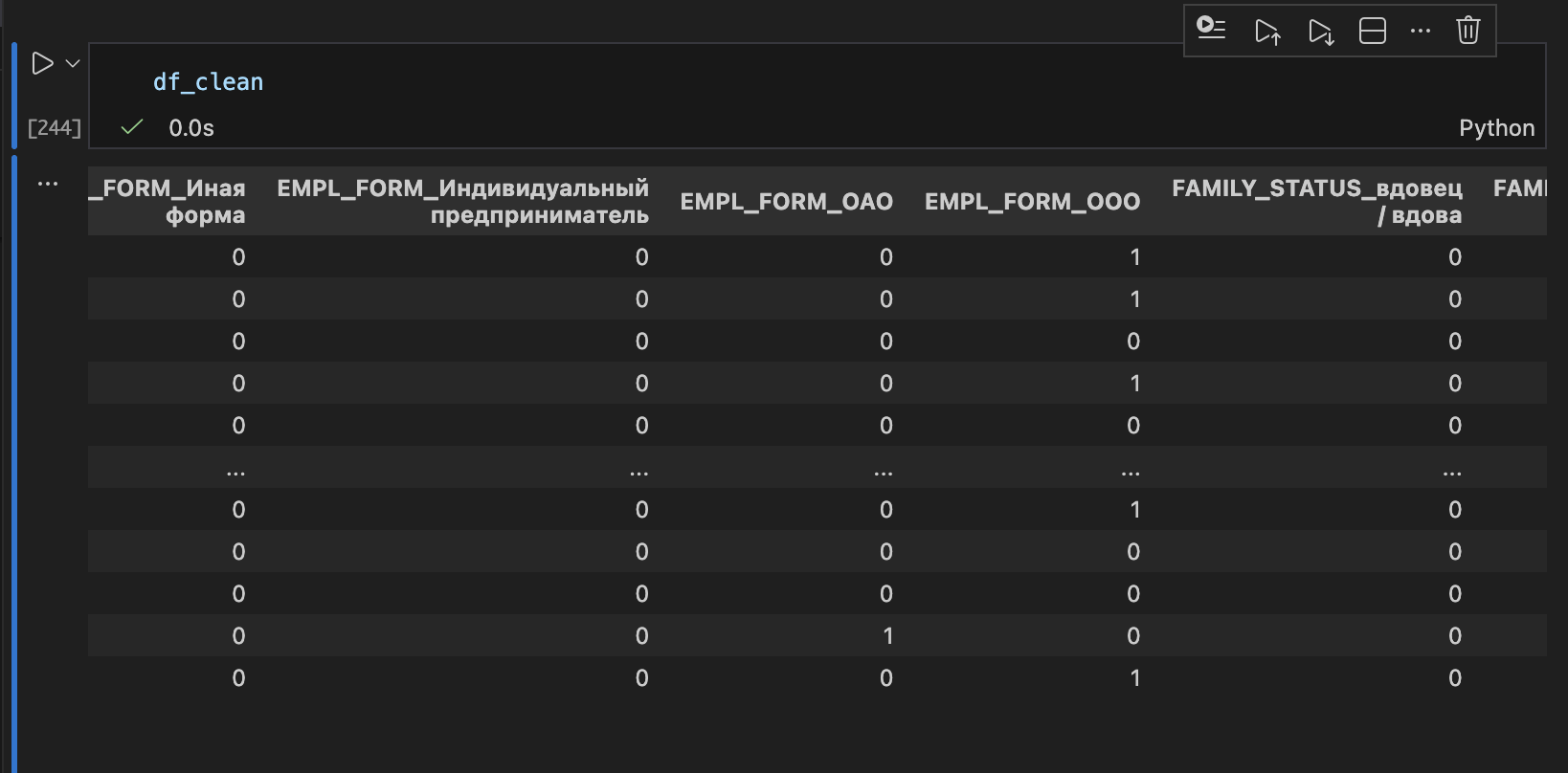
Найдём все большие коэффициенты корреляции:



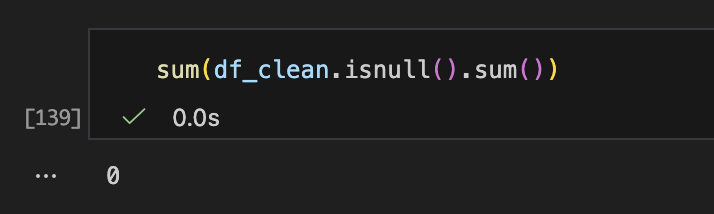
**Итоговые действия с данными:**

1. Удаляем ID и Вариант
2. Меняем единственное значение #ПОЛЕ! в DTI на среднее арифметическое
3. Меняем в BANKACCOUNT\_FLAG значения 2,3 и 4 на 2
4. Меняем \*n.a.\* в EDUCATION на null
5. Меняем в EMPL\_SIZE '>100' на '>=100'
6. Удаляем строки с null значениями в 'thirty\_in\_a\_year', 'sixty\_in\_a\_year', 'ninety\_in\_a\_year', 'thirty\_vintage', 'sixty\_vintage', 'ninety\_vintage' , 'EDUCATION', 'INSURANCE\_FLAG'
7. Заполним NULL значения для количественных переменных медианным значением.
8. Заменим категорию "Высшее/Второе высшее/Ученая степень" на "высшее"
9. Переведём категориальные переменные в целочисленные. Уровень образование переведём в шкалу от 0 до 6 где  
   0 - неполное среднее  
   1 - среднее  
   2 - среднее-специальное  
   3 - незаконченное высшее  
   4 - высшее  
   5 - второе высшее  
   6 - ученая степень  
   чем выше балл - тем выше уровень образования человека.  
   Остальные категориальные !не бинарные! переменные закодируем с помощью one-hot кодирования. Среди них SEX, INCOME\_BASE\_TYPE, CREDIT\_PURPOSE, EMPL\_TYPE, EMPL\_SIZE, EMPL\_PROPERTY, EMPL\_FORM, FAMILY\_STATUS
10. Нормализуем данные  
     

Итого получим: (полная таблица в task1.ipynb)



NULL значений не осталось.



В финальной витрине осталась 3641 строка и 77 переменных. Её мы уже можем использовать для построения сегментации.

# **Часть 2. Сделать минимум два метода сегментации.**

Сделать минимум 2 метода сегментации (Вес задания 30%)

Примеры сегментаций:

* Бизнес-правила;
* Квантили (RFM);
* Кластеризация с учителем (Дерево решений, регрессия, нейросети,
* градиентный бустинг и др.);
* Кластеризация без учителя (Метод K-средних, EM алгоритм, градиентный спуск и др.)

Вне зависимости от выбранного метода сегментации, выделение групп

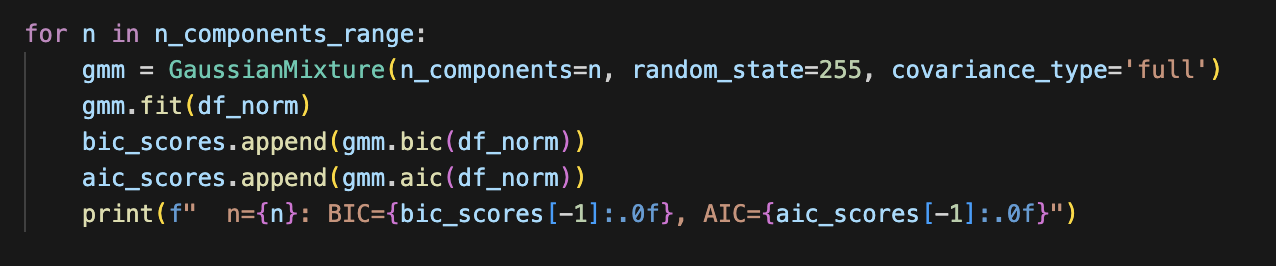
клиентов должно соответствовать следующим условиям:

Внутри сегмента однородность максимальная;

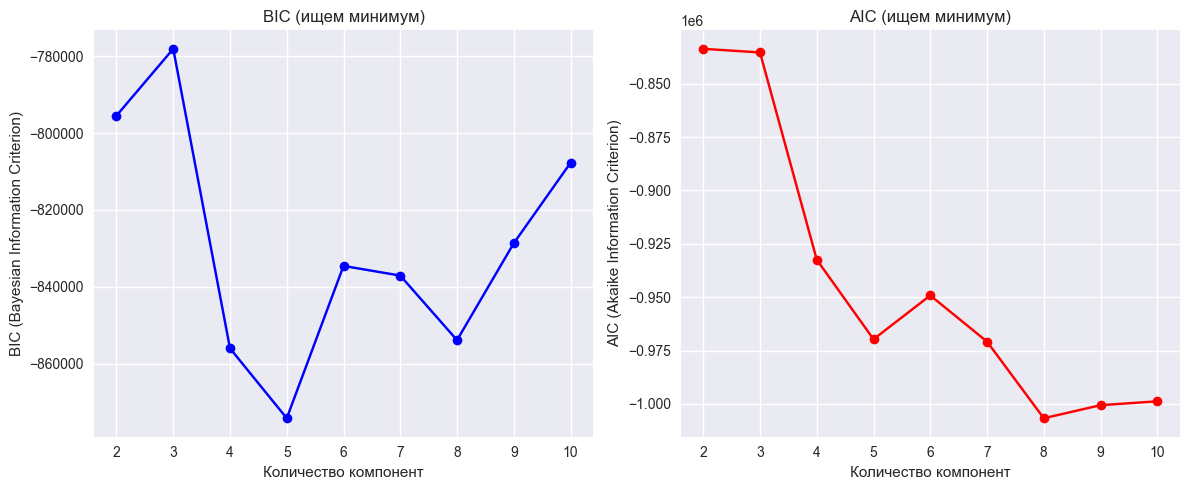
Между сегментами однородность минимальная;

## Часть 2.1. Кластеризация алгоритмом EM

Воспользуемся GaussianMixture и подберём кол-во кластеров

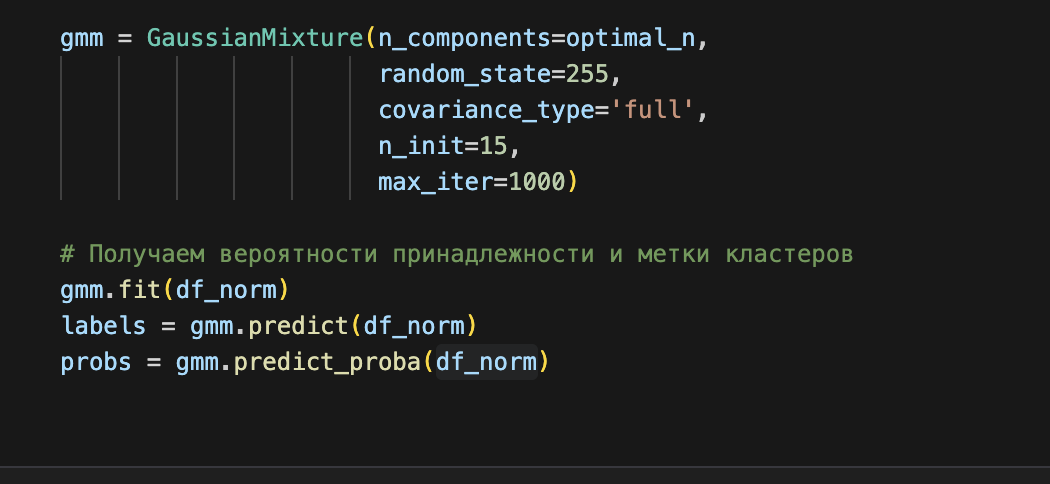


Расчитаем BIC и AIC и выберем в них минимум

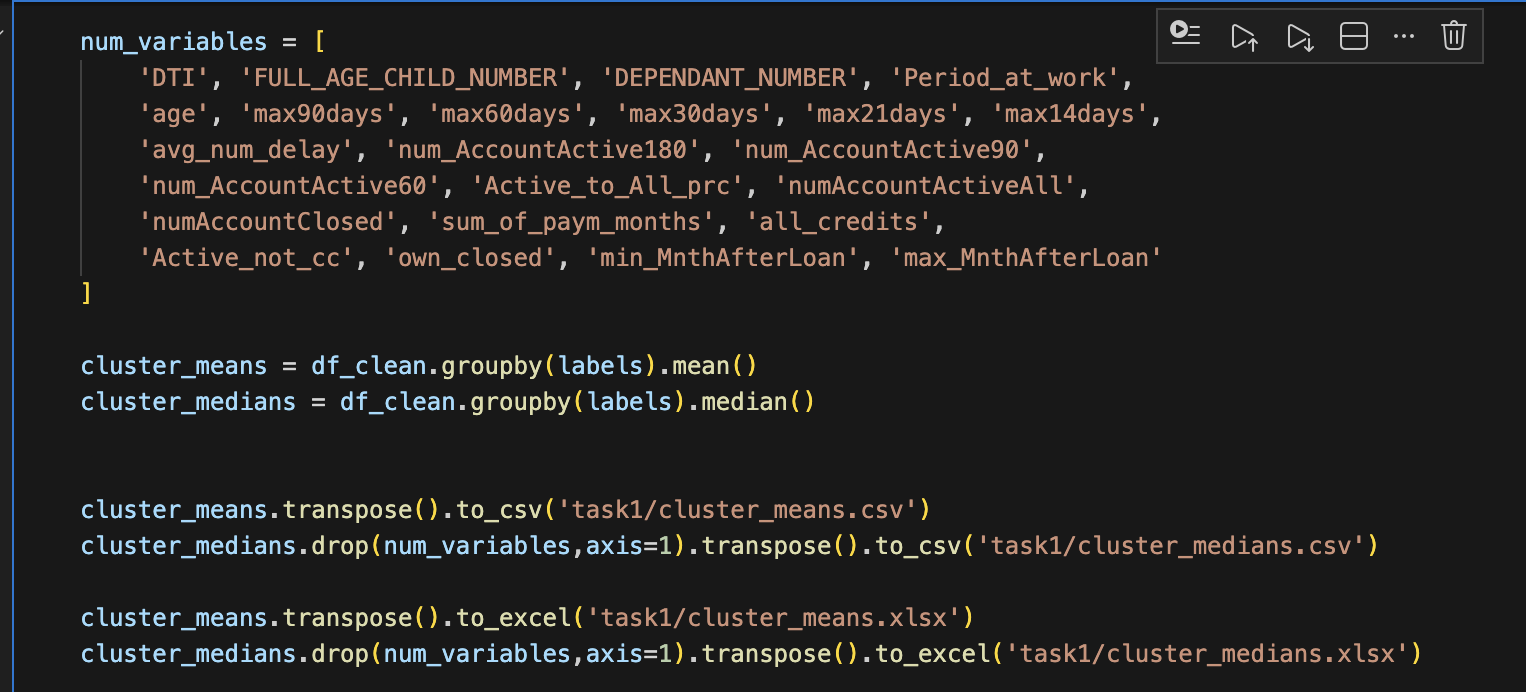


Оптимальным по этим метрикам является число кластеров 6.

Выполним кластеризацию.



Сохраним результаты с разбивкой по кластерам в xlsx файлы



Рассмотрим ключевые признаки в кластерах:

Данные находятся в файлах cluster\_medians.xlsx и cluster\_means.xlsx

Визуалиация t-SNE:



**Кластер 1 - 1248 человек:**

- Женщина 36 лет, без детей

- Зарплата больше 250 тысяч

- Работает в ООО'шке, скорее всего в сфере торговли по профессии специалист

- Семейное положение: замужем либо никогда не была замужем

- Имеет высшее образование

- Сумма платежей в месяц 76 тысяч

- Имеет задолженности но редко более чем 60 дней

**Кластер 2 (менеджер-специалист) - 783 человек:**

- Зарплата менее 50ти тысяч

- Замужем и скорее всего есть ребёнок

- Род занятости

- Цель кредита - Ремонт, чаще других кластеров кредит под залог залог

- Имеет задолженности но редко более чем 60 дней

- Платежи в месяц - 90 тысяч, живёт в минус

**Кластер 3 - 1167 человек:**

- Мужчина 35 лет, женат, вероятно есть 1 ребёнок

- Зарплата более 250 тысяч

- Цель кредита - Ремонт или покупка автомобиля

- Специалист или менеджер с сфере торговли

**Кластер 4 (должник) - 195 человек:**

- Группа людей который очень редко выплачивают кредиты вовремя

- в 3-6 раз чаще чем у других кластеров имеет задолженность в течении 90 дней

- Работает в сфере торговле

**Кластер 5 (гос работник) - 248 человек:**

- с шансом 80% работают в неизвестной сфере

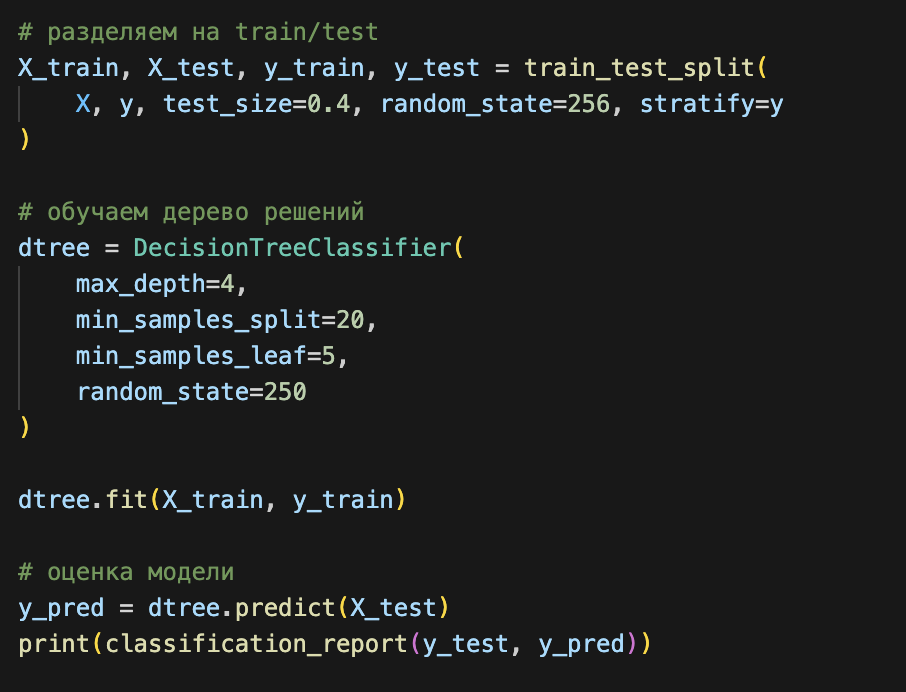
- Специалист с зарплатой >250

- Работает в государственном предприятии

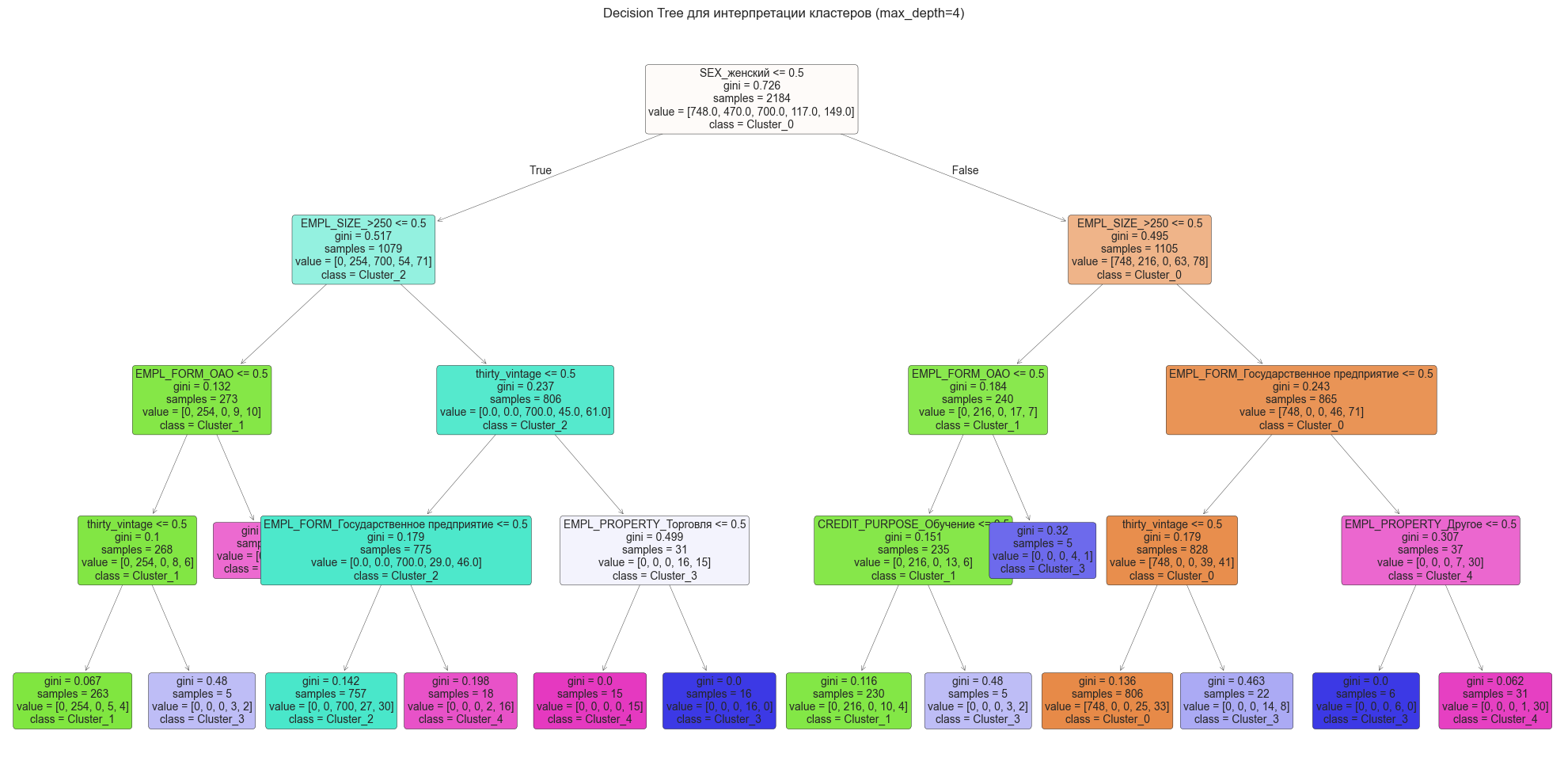
- Подтверждение дохода 2НДФЛ

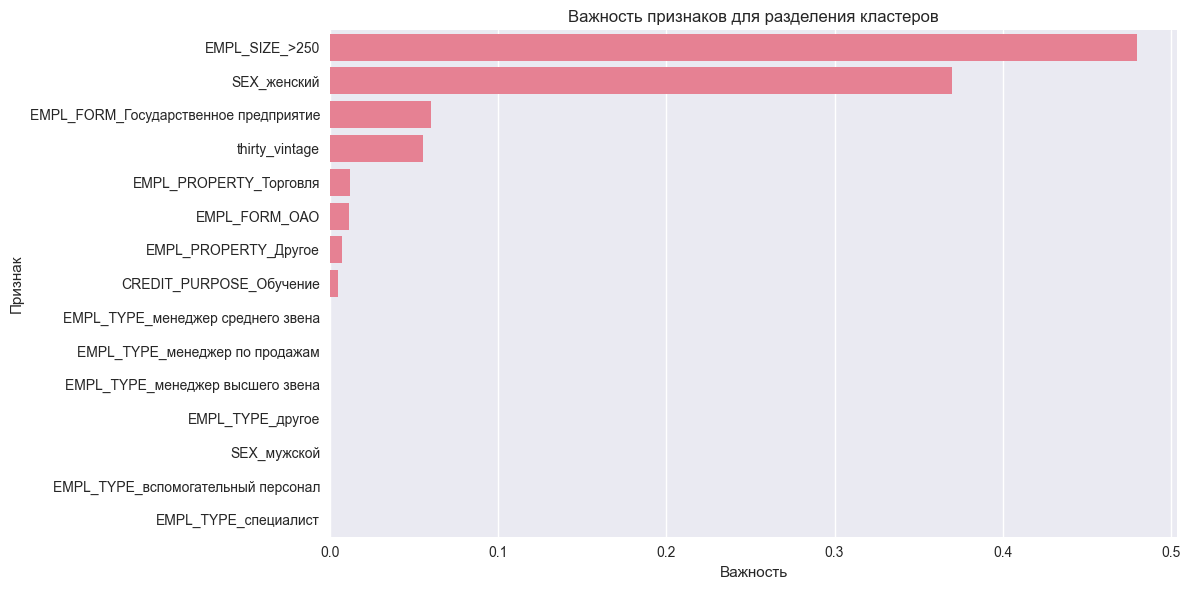
## Часть 2.2 Кластеризация методом дерева решений

Возьмём тот же GMM, из прошлого метода  


Обучим модель на тестовой выборке и оценим её  


Визуализируем модель и оценим признаки по важности для разделения





Итак можно выделить 5 ключевых признаков для разделения:

1) Зарплата, в нашей выборке осталось только две категории >250 и <50

2) Пол

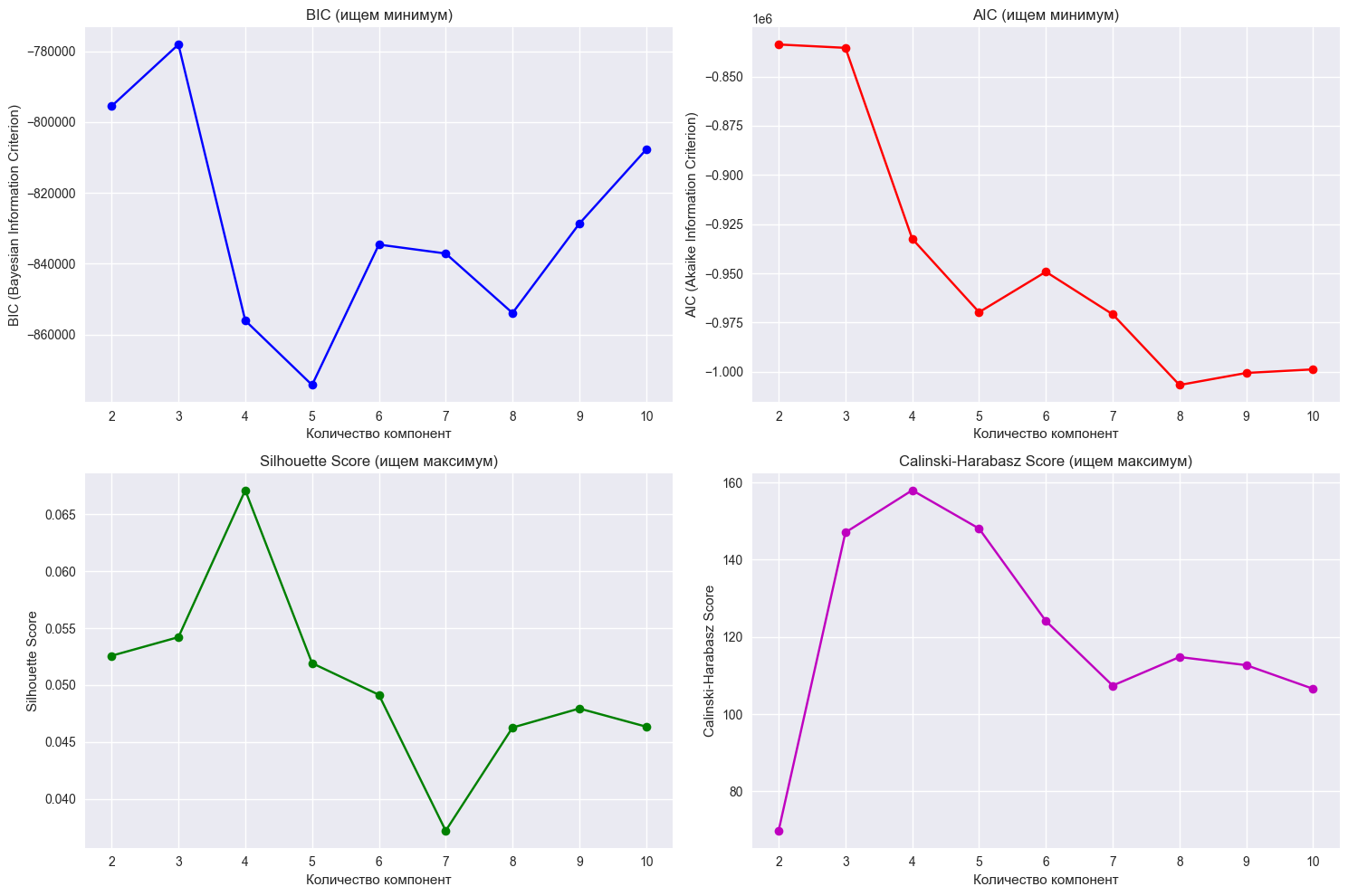
3) Форма предприятия - чаще всего ООО или Государственное

4) Наличие просрочки в 30 дней

5) Сфера работы

# Часть 3 - Оценка по определению количества кластеров

1) Кластеризация без учителя



Выбор был сделан в пользу 5 кластеров по метрикам BIC, CHS и AIC, метод силуэта дал не особо значимые результаты.

Pie Chart:  


2) Кластеризация с учителем

precision recall f1-score support

0 0.93 1.00 0.96 500

1 0.95 1.00 0.97 313

2 0.93 1.00 0.96 467

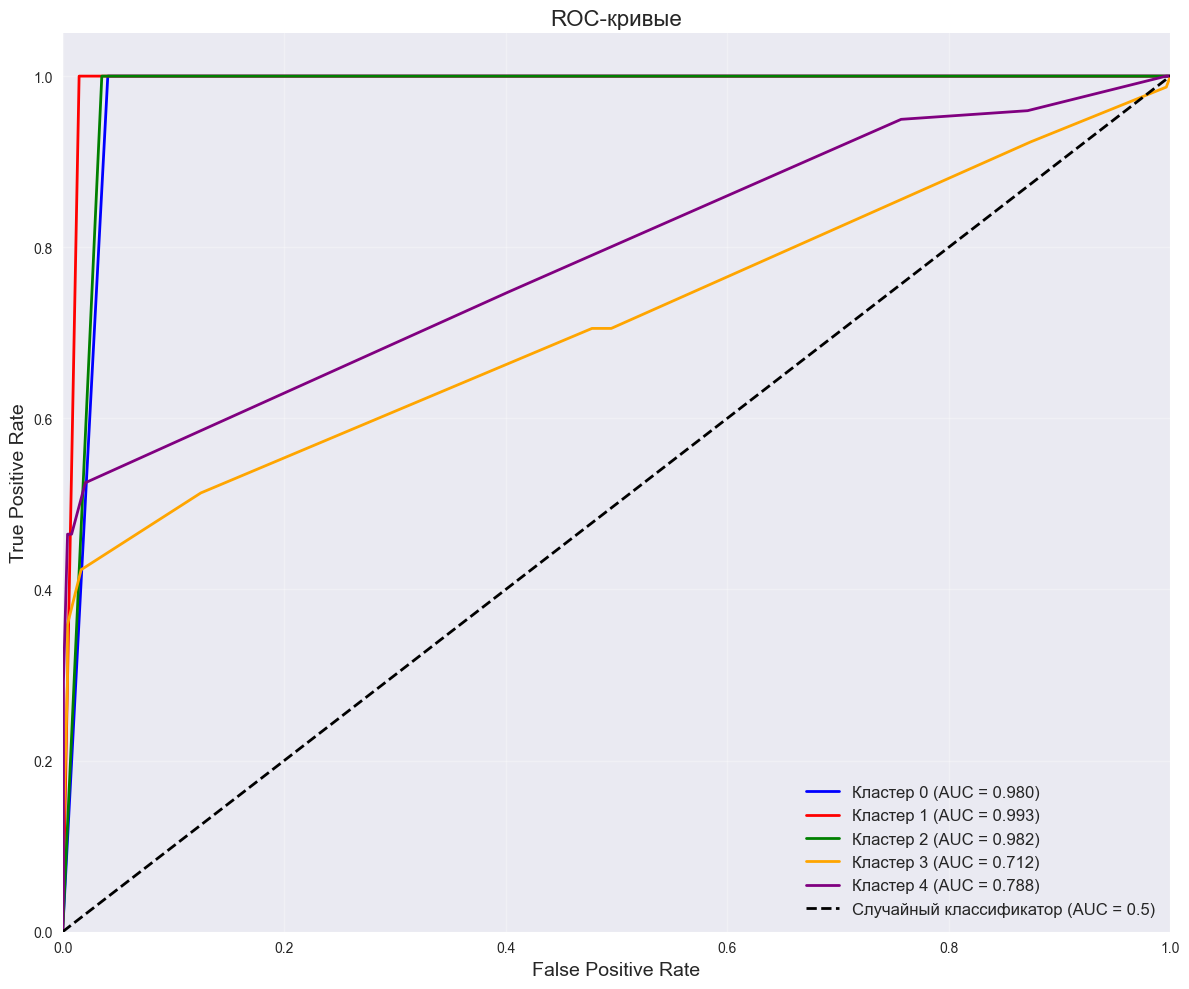
3 0.82 0.35 0.49 78

4 0.87 0.46 0.61 99

accuracy 0.93 1457

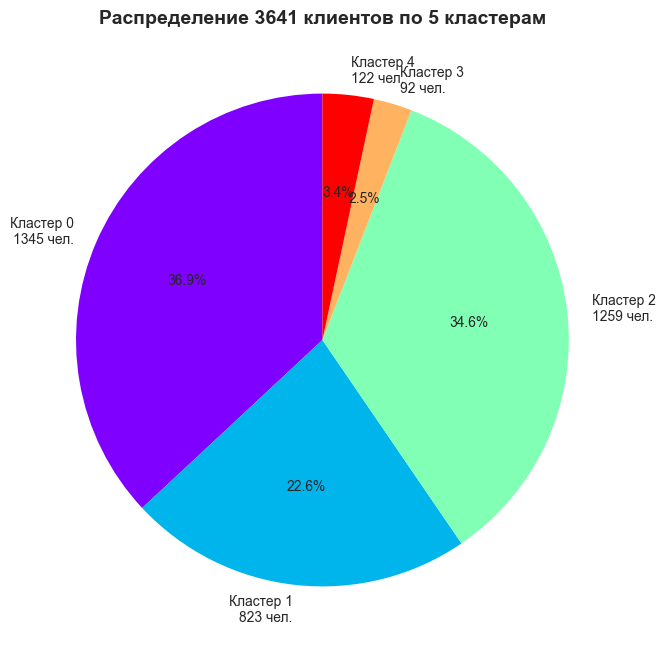
macro avg 0.90 0.76 0.80 1457

weighted avg 0.92 0.93 0.92 1457



Видим хороший результат для кластеров 0, 1, 2 и приемлемый для 3,4 скорее всего дерево путает людей между этими кластерами, кроме того оба этих кластера имеют сильно меньше людей (195 и 248)

PieChart для полного предсказания дерева решения по начальным данным:



# Часть 4. Профили клиентов

**Кластер 1 - 1248 человек:**

- Женщина 36 лет, без детей

- Зарплата больше 250 тысяч

- Работает в ООО'шке, скорее всего в сфере торговли по профессии специалист

- Семейное положение: замужем либо никогда не была замужем

- Имеет высшее образование

- Сумма платежей в месяц 76 тысяч

- Имеет задолженности но редко более чем 60 дней

**Кластер 2 (менеджер-специалист) - 783 человек:**

- Зарплата менее 50ти тысяч

- Замужем и скорее всего есть ребёнок

- Род занятости

- Цель кредита - Ремонт, чаще других кластеров кредит под залог залог

- Имеет задолженности но редко более чем 60 дней

- Платежи в месяц - 90 тысяч, живёт в минус

**Кластер 3 - 1167 человек:**

- Мужчина 35 лет, женат, вероятно есть 1 ребёнок

- Зарплата более 250 тысяч

- Цель кредита - Ремонт или покупка автомобиля

- Специалист или менеджер с сфере торговли

**Кластер 4 (должник) - 195 человек:**

- Группа людей который очень редко выплачивают кредиты вовремя

- в 3-6 раз чаще чем у других кластеров имеет задолженность в течении 90 дней

- Работает в сфере торговле

**Кластер 5 (гос работник) - 248 человек:**

- с шансом 80% работают в неизвестной сфере

- Специалист с зарплатой >250

- Работает в государственном предприятии

- Подтверждение дохода 2НДФЛ