# Message Features

['date', 'sender\_id', 'message', 'char\_count', 'word\_count', 'has\_photo',

'has\_video', 'has\_audio', 'has\_document', 'has\_is\_reply',

'has\_is\_sticker', 'sentiment', 'is\_question', 'is\_admin\_post',

'cluster', 'week', 'month', 'day\_of\_week', 'hour']

Сколько прошло времени перед предыдущим сообщением

Сколько прошло времени перед предыдущим сообщением этого же пользователя

# Анализ сообщений

{'Количество сообщений': 53938,

'Количество слов': 652213,

'Количество фото': 3350,

'Количество видео': 81,

'Количество аудио': 16,

'Количество документов': 535,

'Количество стикеров': 157,

'Количество вопросов': 12113,

'Количество ответов': 20841,

'Количество сообщений администратора': 321,

'Среднее количество слов в сообщении': 12.091901813192925,

'Среднее количество букв в сообщении': 81.65434387630242,

'Сколько в среднем фото на сообщение': 0.06210834662019356,

'Сколько в среднем видео на сообщение': 0.0015017242018613964,

'Сколько в среднем аудио на сообщение': 0.00029663687938002894,

'Сколько в среднем документов на сообщение': 0.009918795654269718,

'Сколько в среднем стикеров на сообщение': 0.002910749378916534,

'Сколько в среднем вопросов на сообщение': 0.22457265749564315,

'Сколько в среднем ответов на сообщение': 0.38638807519744894,

'Сколько в среднем сообщений администратора': 0.00595127739256183,

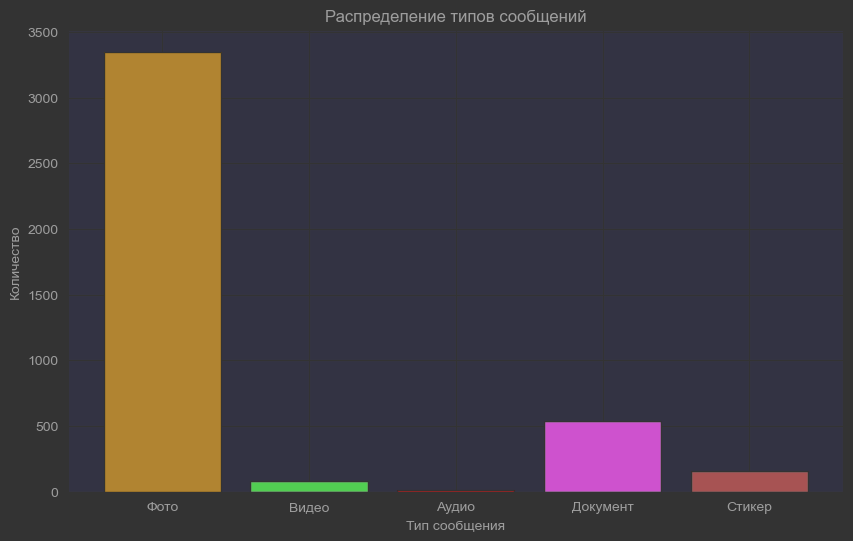
'Соотношение вопросов и ответов': 0.5812101146777986}

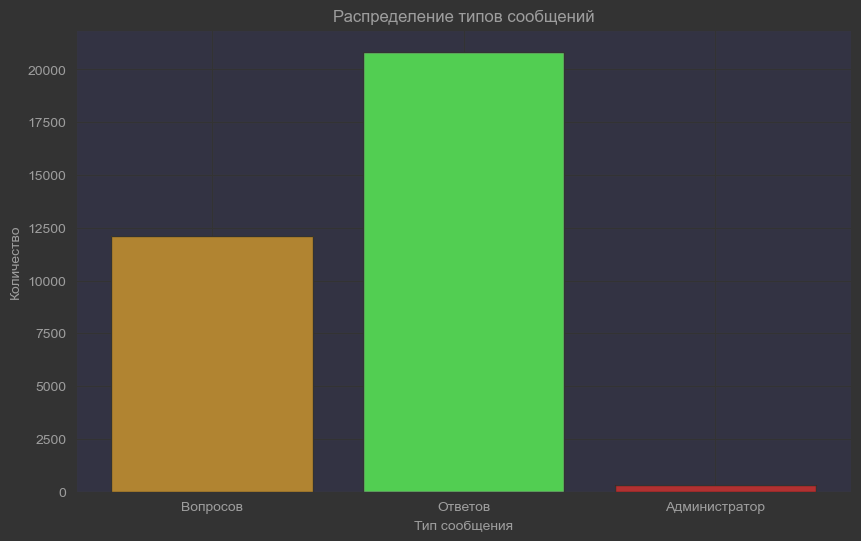
Среднее время ответа (как посчитать ?)

## Распределения

Распределение типов сообщений (Текст, Видео, Фото , Аудио, документ, стикер)

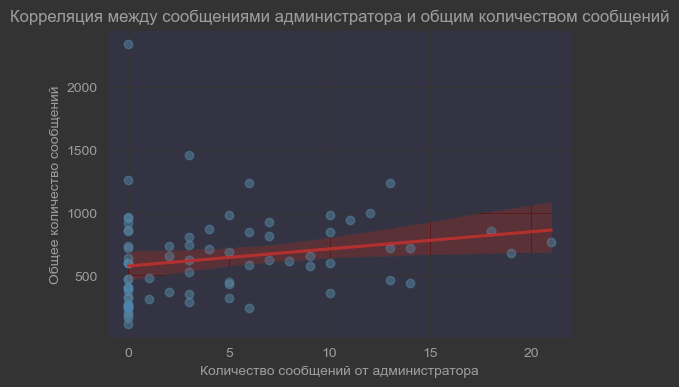
{'Фото': 3350, 'Видео': 81, 'Аудио': 16, 'Документ': 535, 'Стикер': 157}





## Корреляционный Анализ

Корреляция между количеством сообщений от администратора и общим количеством сообщений: 0.20433907461498446 (Месячные данные)



## Сентимент Анализ

Сентимент анализа:

sentiment

1 star 17446

5 stars 11845

3 stars 8200

neutral 6999

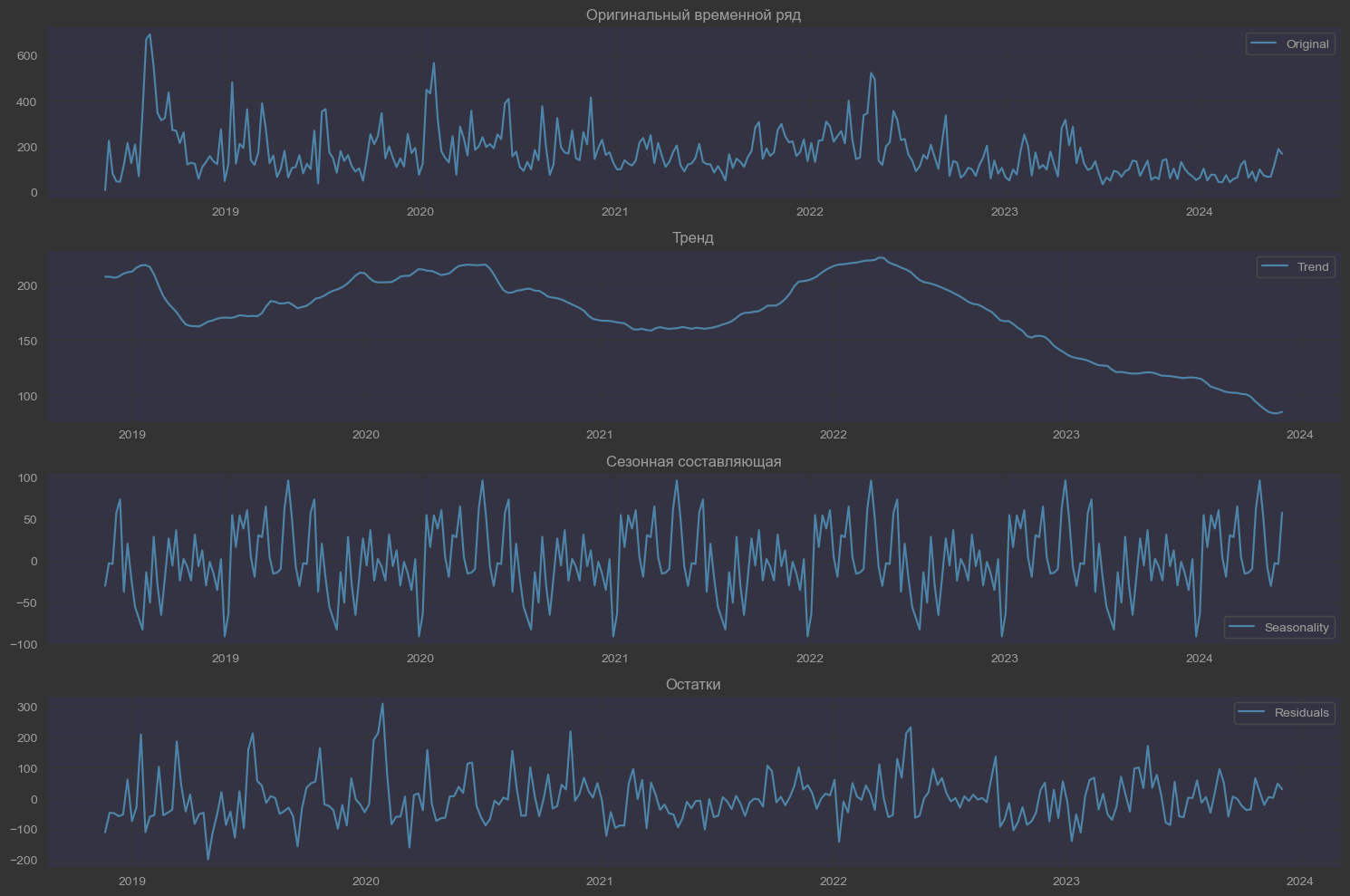
4 stars 5283

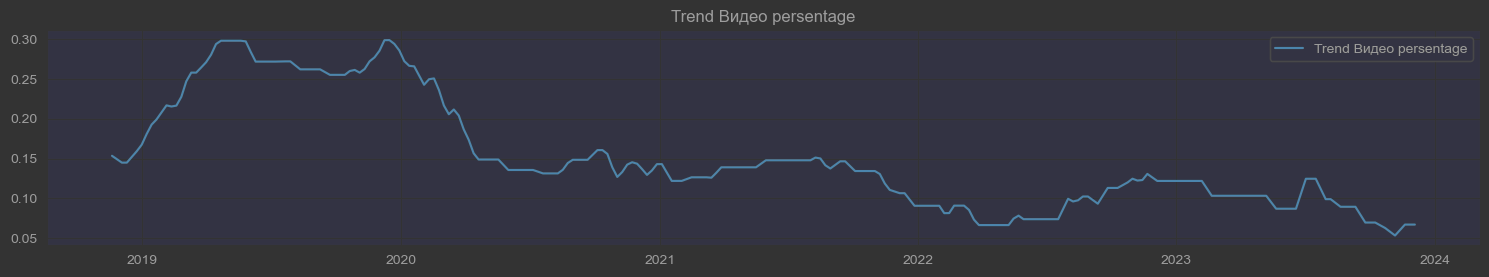
2 stars 4165

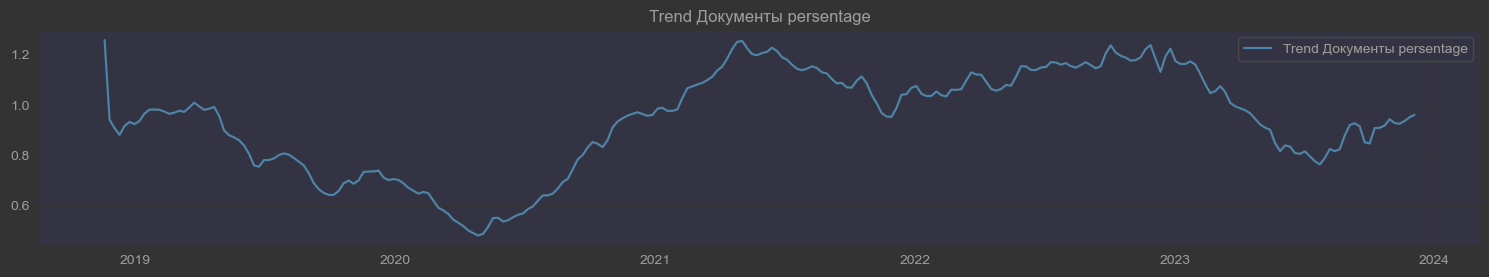
Самые популярные слова: [('aimylogic', 3168), ('бота', 2986), ('спасибо', 2936), ('бот', 2915), ('https', 2609), ('будет', 2540), ('com', 2442), ('добрый', 2393), ('день', 2206), ('может', 2188)]

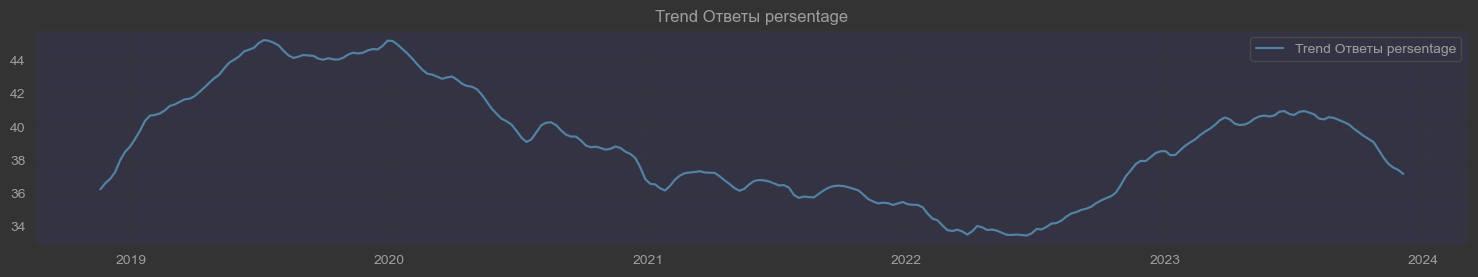
## Анализ временных рядов

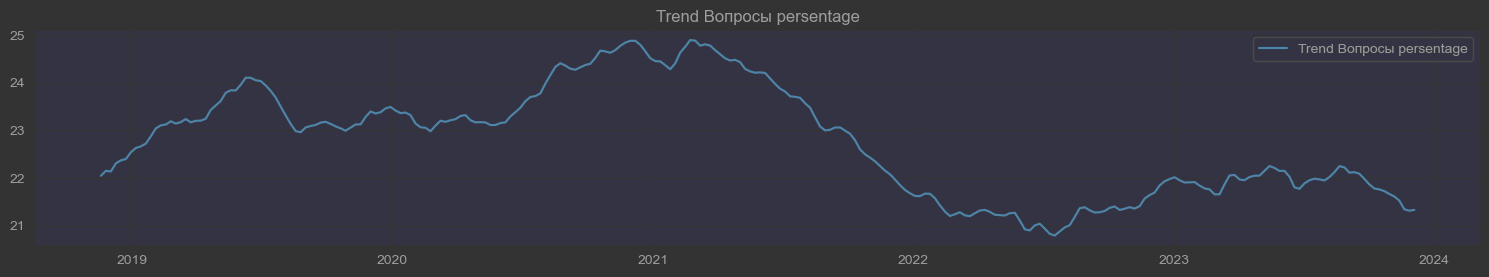
### Декомпозиция временных рядов











# User Features

['sender\_id', 'message\_count', 'avg\_char\_count', 'total\_char\_count',

'avg\_word\_count', 'total\_word\_count', 'avg\_photo', 'total\_photo',

'avg\_video', 'total\_video', 'avg\_audio', 'total\_audio', 'avg\_document',

'total\_document', 'lifetime\_days', 'first\_message\_date',

'last\_message\_date', 'avg\_is\_question', 'total\_is\_question',

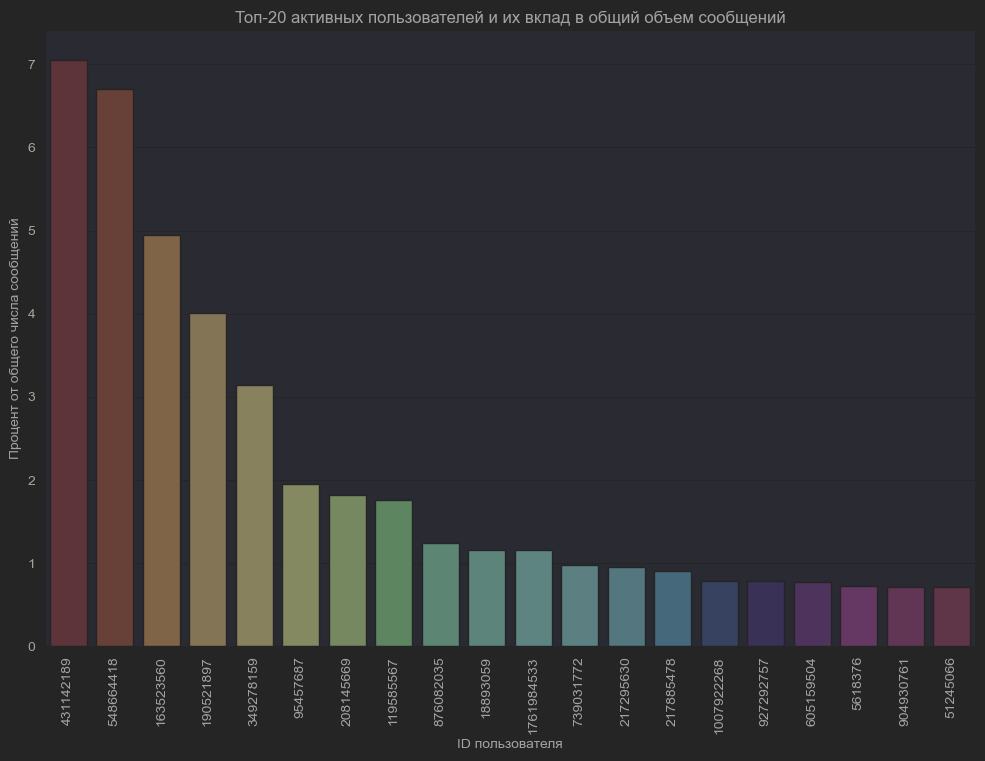
'avg\_is\_reply', 'total\_is\_reply', 'is\_active', 'most\_active\_hour',

'neutral\_messages', '1star\_messages', '2star\_messages',

'3star\_messages', '4star\_messages', '5star\_messages', 'cluster']

# Анализ пользователей

## Распределения

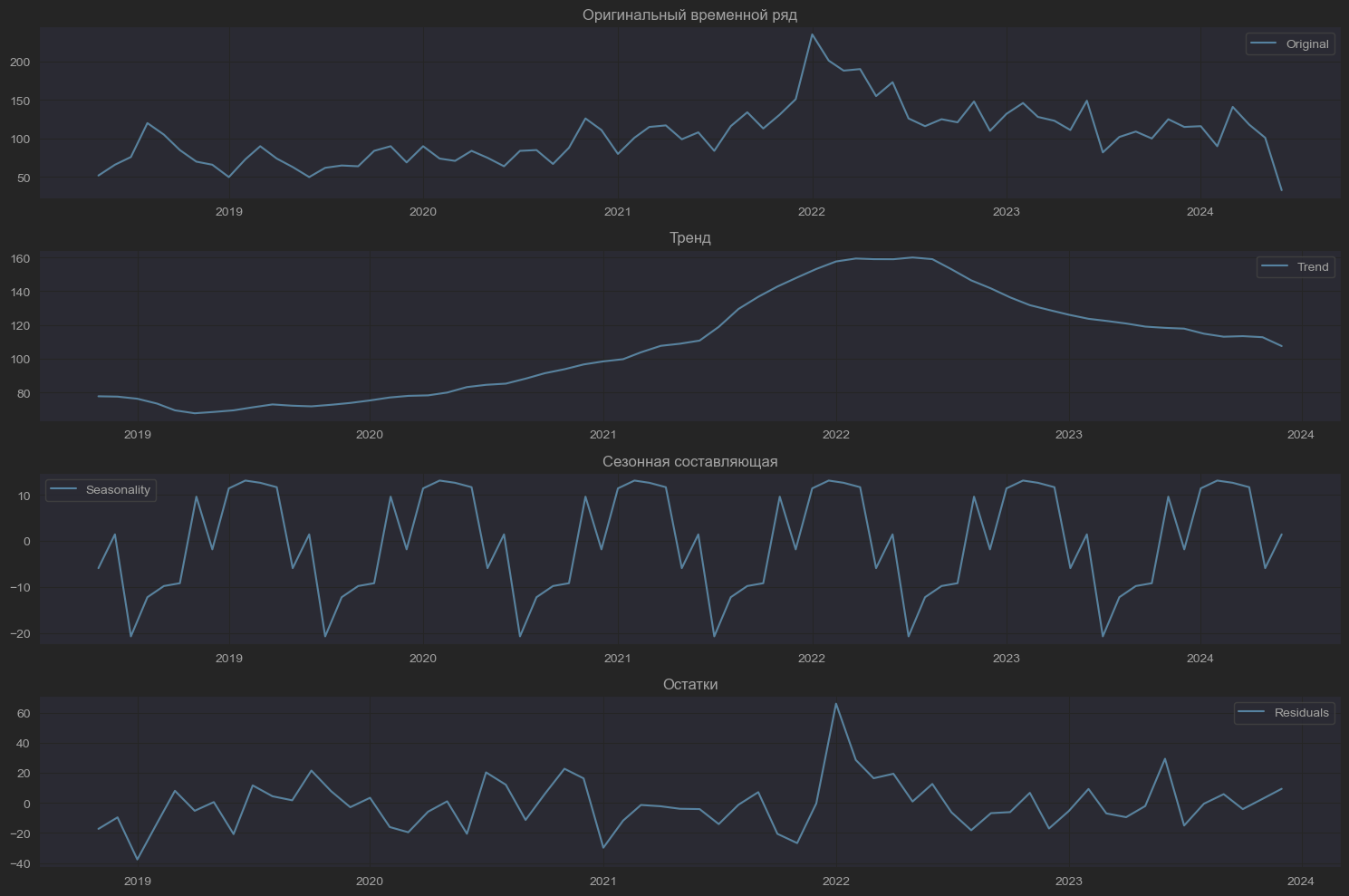


Процент пользователей, которые создают 90% сообщений: 20.84%

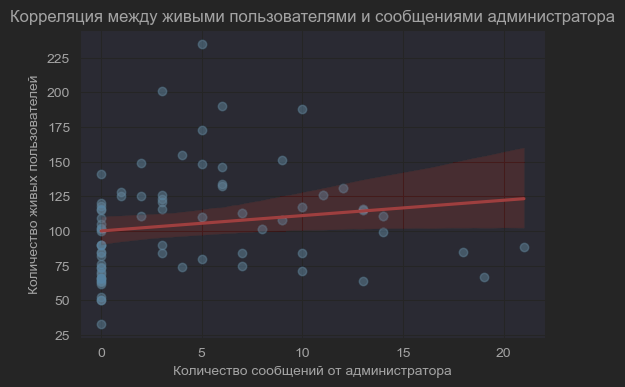
## Анализ временных рядов

### Декомпозиция

Количество живых пользователей



## Корреляционный анализ

Корреляция между количеством живых пользователей и количеством сообщений администратора: 0.15614268953629326

## Кластеризация

| cluster | -1 | 0 |
| --- | --- | --- |
| sender\_id | 2003 | 3025 |
| message\_count | 25.288567 | 1.072397 |
| avg\_char\_count | 72.537942 | 0.399364 |
| total\_char\_count | 2192.873689 | 1.228430 |
| avg\_word\_count | 10.687749 | 0.066754 |
| total\_word\_count | 324.796306 | 0.205620 |
| avg\_photo | 0.044466 | 0.000000 |
| total\_photo | 1.670994 | 0.000000 |
| avg\_video | 0.001891 | 0.000000 |
| total\_video | 0.040439 | 0.000000 |
| avg\_audio | 0.000086 | 0.000000 |
| total\_audio | 0.007988 | 0.000000 |
| avg\_document | 0.009743 | 0.000000 |
| total\_document | 0.2666 | 0.0000 |
| lifetime\_days | 123.794808 | 0.528595 |
| avg\_is\_question | 0.26928 | 0.00000 |
| total\_is\_question | 6.047429 | 0.000000 |
| avg\_is\_reply | 0.196553 | 0.000000 |
| total\_is\_reply | 10.391912 | 0.000000 |
| is\_active | 0.158263 | 0.000000 |
| most\_active\_hour | 11.932102 | 12.394050 |
| neutral\_messages | 1.968048 | 1.009917 |
| 1star\_messages | 8.686970 | 0.012893 |
| 2star\_messages | 2.076385 | 0.000992 |
| 3star\_messages | 4.079880 | 0.008926 |
| 4star\_messages | 2.628058 | 0.004628 |
| 5star\_messages | 5.849226 | 0.035041 |

# Выводы

### Инсайты по анализу сообщений

#### Общее количество сообщений и их распределение

* Общее количество сообщений: 53,938
* Общее количество слов: 652,213
* Среднее количество слов в сообщении: 12.09
* Среднее количество букв в сообщении: 81.65

#### Типы сообщений

* Фото: 3,350 (6.21% от общего количества сообщений)
* Видео: 81 (0.15% от общего количества сообщений)
* Аудио: 16 (0.03% от общего количества сообщений)
* Документы: 535 (0.99% от общего количества сообщений)
* Стикеры: 157 (0.29% от общего количества сообщений)

#### Вопросы и ответы

* Количество вопросов: 12,113 (22.46% от общего количества сообщений)
* Количество ответов: 20,841 (38.64% от общего количества сообщений)
* Соотношение вопросов и ответов: 0.58

#### Администраторские сообщения

* Количество сообщений администратора: 321 (0.60% от общего количества сообщений)

### Тренды и корреляции

1. Тренды по типам сообщений:
   * За последний год процент фото от общего количества сообщений вырос в два раза.
   * Количество видео и аудио сообщений осталось неизменным.
   * Количество документов незначительно снизилось.
   * Процент ответов также вырос за последний год и вероятно коррелирует с количеством отправленных фото.
   * Вопросы показали незначительный рост.
2. Общее количество сообщений:
   * Общее количество сообщений падает.
   * Наблюдается ярко выраженная сезонность, с пиками в четвертый месяц и просадками в двенадцатый.
   * Общий тренд на количество сообщений показывает ярко выраженное падение.
   * Количество активных пользователей (хотя бы одно сообщение за месяц) показывает меньшее падение.
3. Корреляции:
   * Корреляция между количеством сообщений и постами новостей из аффилированного канала составляет 20%.
   * Корреляция между количеством живых пользователей и количеством сообщений постов от аффилированного канала составляет 15%.
4. Активность по дням недели и времени суток:
   * Наименее активные дни недели: суббота и воскресенье.
   * Наиболее активные дни недели: среда и четверг.
   * Пик активности приходится на период с 7 утра до 15 часов.
5. Процент пользователей, создающих 90% сообщений:
   * 20.84% пользователей создают 90% сообщений.

### Кластеризация пользователей

* Метод: DBSCAN
* Результаты:
  + Первая группа: 3,000 пользователей, сделавших чуть больше одного сообщения. Среднее время жизни пользователя составило полдня.
  + Вторая группа: 2,000 пользователей, более активных, с в среднем 25 сообщениями и временем жизни в 123 дня (время от первого до последнего сообщения).

### Сентимент-анализ

* Метод: BERTA
* Результаты:
  + 1 звезда: 17,446 сообщений
  + 5 звезд: 11,845 сообщений
  + 3 звезды: 8,200 сообщений
  + Нейтрально: 6,999 сообщений
  + 4 звезды: 5,283 сообщений
  + 2 звезды: 4,165 сообщений

### Самые популярные слова

* Топ-10 популярных слов:
  + aimylogic (3,168 упоминаний)
  + бота (2,986 упоминаний)
  + спасибо (2,936 упоминаний)
  + бот (2,915 упоминаний)
  + https (2,609 упоминаний)
  + будет (2,540 упоминаний)
  + com (2,442 упоминаний)
  + добрый (2,393 упоминаний)
  + день (2,206 упоминаний)
  + может (2,188 упоминаний)

### Выводы и рекомендации

1. Активность пользователей:
   * Необходимо разработать стратегию для увеличения активности пользователей, особенно в периоды низкой активности (декабрь) и в выходные дни.
   * Разработать специальные акции или мероприятия, чтобы стимулировать активность в эти периоды.
2. Типы контента:
   * Учитывая рост популярности фото, стоит продолжать поощрять пользователей делиться фотографиями.
   * Рассмотреть возможность интеграции новых типов контента или улучшения существующих, чтобы поддержать интерес пользователей.
3. Сентимент-анализ:
   * Несмотря на сомнительные результаты BERTA, анализ настроений показывает, что большинство сообщений имеют положительный или нейтральный тон. Это может быть использовано для улучшения пользовательского опыта и поддержки позитивной атмосферы в сообществе.
4. Кластеризация пользователей:
   * Для первой группы пользователей (менее активных) можно разработать стратегии вовлечения, такие как персонализированные уведомления или предложения.
   * Для второй группы (более активных) можно предложить дополнительные возможности и привилегии, чтобы поддерживать их активность.
5. Корреляции и сезонность:
   * Учитывая корреляцию между постами новостей и количеством сообщений, можно рассмотреть возможность увеличения частоты постов в аффилированном канале в периоды низкой активности, чтобы стимулировать обсуждения.
   * Анализ сезонности может помочь в планировании маркетинговых и коммуникационных кампаний, чтобы максимизировать вовлеченность пользователей.

# приложение

