

Итоговый проект по программе “Основы анализа данных в научной деятельности”

Цифровые кафедры Казанского федерального университета

# Обзор научных исследований на тему “Новая коронавирусная инфекция (COVID-19)” за период 2020–2021 г.

Подготовили ординаторы ИФМИБ КФУ:

Рамазанова Миляуша Илдаровна

Ефимова Диляра Маратовна

Нуриягдыева Эмине

Гайнутдинова Аделина Рустемовна

Шевченко Роман Васильевич

Минязева Ирина Салаватовна

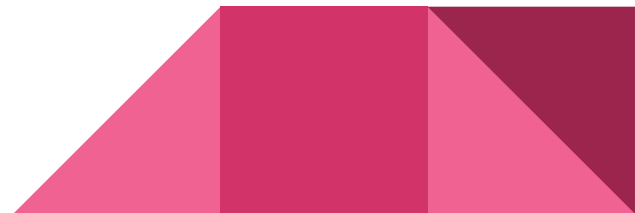
# Актуальность

## Пандемия COVID-19

**Резкое увеличение количества научных публикаций**

**Необходимость оценки качества и значимости новых исследований**

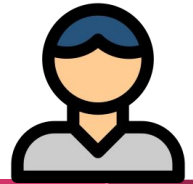
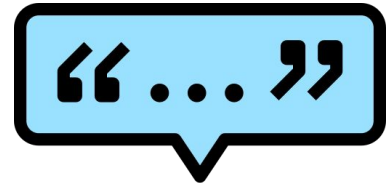
**Понимание факторов, влияющих на "успех" публикации, помогает исследователям, редакторам и аналитикам выявлять важные научные тренды и принимать обоснованные решения.**




**Ключевая проблема:** Существуют тысячи научных статей по теме COVID-19, но лишь часть из них получает широкое признание в научном сообществе.

Какие характеристики публикации могут влиять на её успех?

**Цель:** проанализировать взаимосвязи между признаками публикаций и их успешностью, используя открытый датасет CORD-19 с помощью средств визуального анализа.



## Задачи проекта:

- Подготовить и очистить данные о публикациях.
  - Создать бинарную переменную успешности публикации.
  - Оценить уровень кооперации между авторами.
  - Построить визуализации распределений, корреляций и временных трендов.
  - Выявить связи между кооперацией, временем публикации и цитируемостью.
  - Сформулировать **выводы** о потенциальных факторах успеха публикаций.
- 

# Источники данных

Набор данных исследовательских работ по COVID-19  
<https://www.kaggle.com/draaslan/covid19-research-papers-dataset>

Каждая статья имеет следующие столбцы данных:

- PubMed ID
- DOI
- Journal Name
- Journal Country
- Paper Title
- Authors
- Abstract
- Publication Date
- Citation Count

Всего он содержит 165 000 статей .

Все статьи со словом «COVID-19», опубликованные до сентября 2021 года, были включены в набор данных.

kaggle

Абдуссамет Аслан,  
доктор медицины,  
Анкара, Турция



# Подготовка данных (что было сделано, какие проблемы и каким образом решили)

Работа выполнена в программе RStudio. Были загружены основные библиотеки: tidyverse, lubridate, readr, patchwork, ggcorrplot.

1. Импортирован датасет COVID-19 с Kaggle.
2. Исследованы ключевые поля: авторы, дата публикации, количество цитирований, журналы и др.
3. Удалены дубликаты и строки с пропущенными значениями в важных колонках; форматированы даты.
4. Создание новых переменных:
  - collaboration\_level — количество авторов в статье.
  - month\_year\_num — порядковый номер месяца публикации.
  - successful — бинарная переменная: 1, если статья выше медианы по числу цитирований.
5. Построены гистограммы, боксплоты, плотности распределений, временные тренды и корреляционные матрицы.

6. Проанализированы распределения цитирований, уровни кооперации и временная динамика публикаций.

7. Выявлены слабые, но интересные связи между числом авторов, временем публикации и успешностью статей.

8. Отмечены публикации с экстремально высокими цитированиями, которые сильно влияют на распределение.

9. Проводилось уточнение переменных, пересчёт успешности и пересмотр порогов.

10. Все шаги задокументированы; подготовлены визуализации и выводы, готовые для презентации проекта.



# Первичный анализ и визуализация (графики, основные наблюдения)

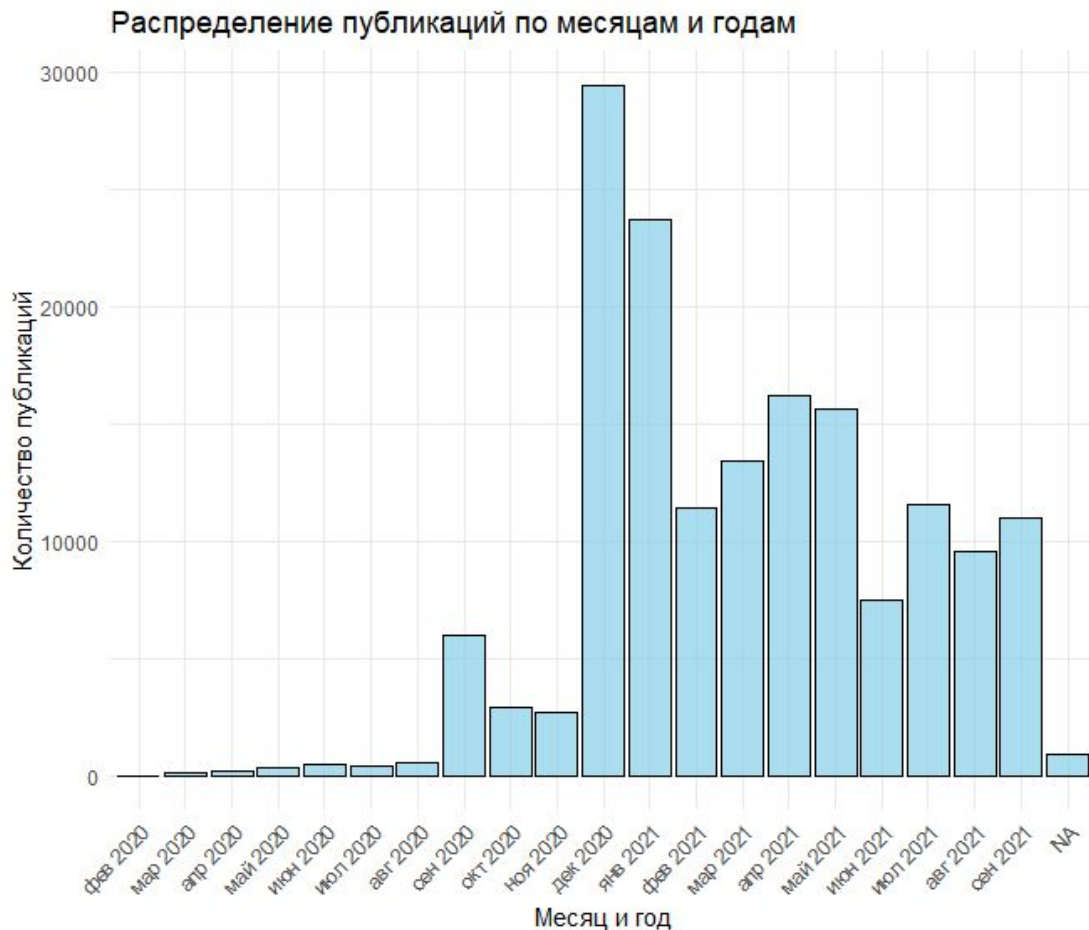
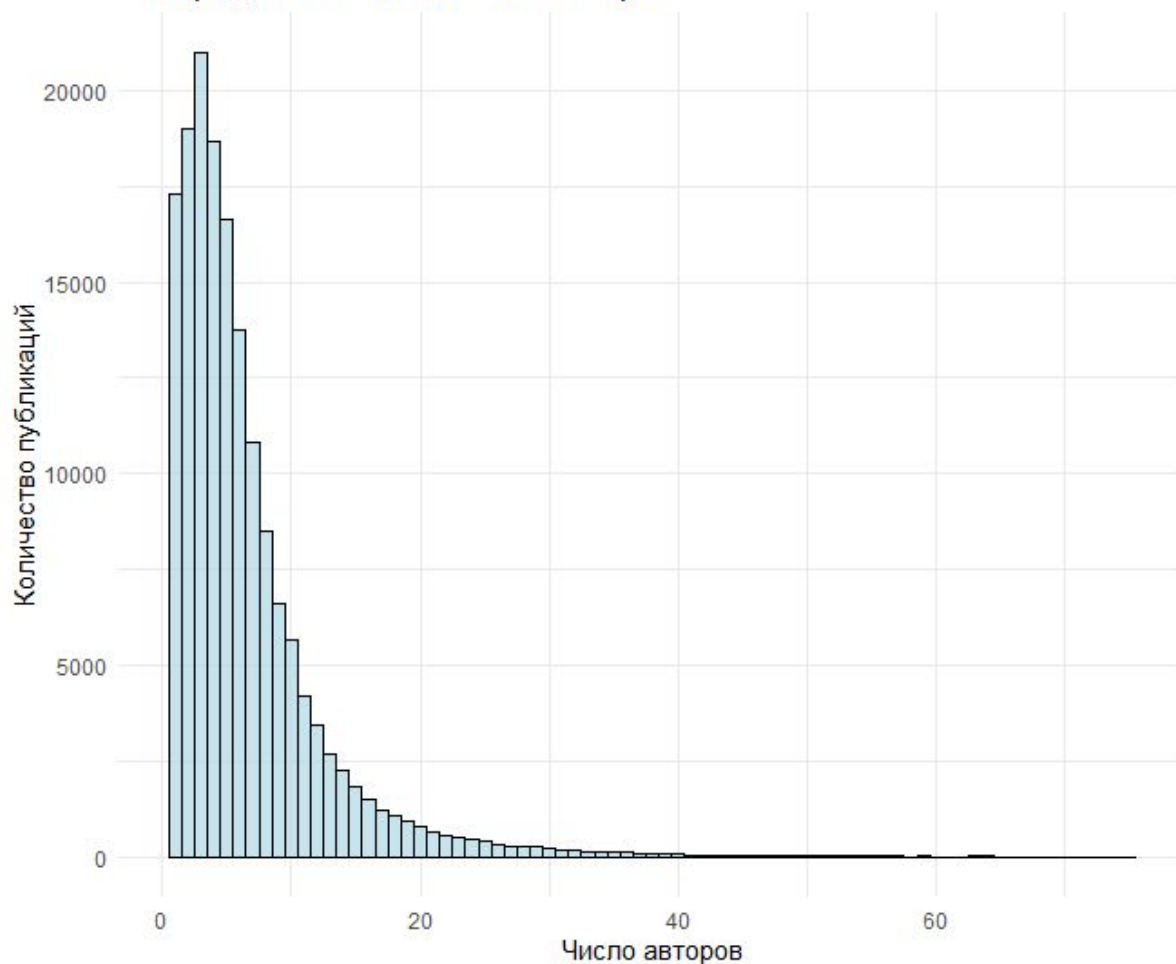


График публикаций по месяцам и годам. Т.к. все публикации 2020 и 2021 года, было принято решение дополнительно использовать месяцы для лучшей визуализации.

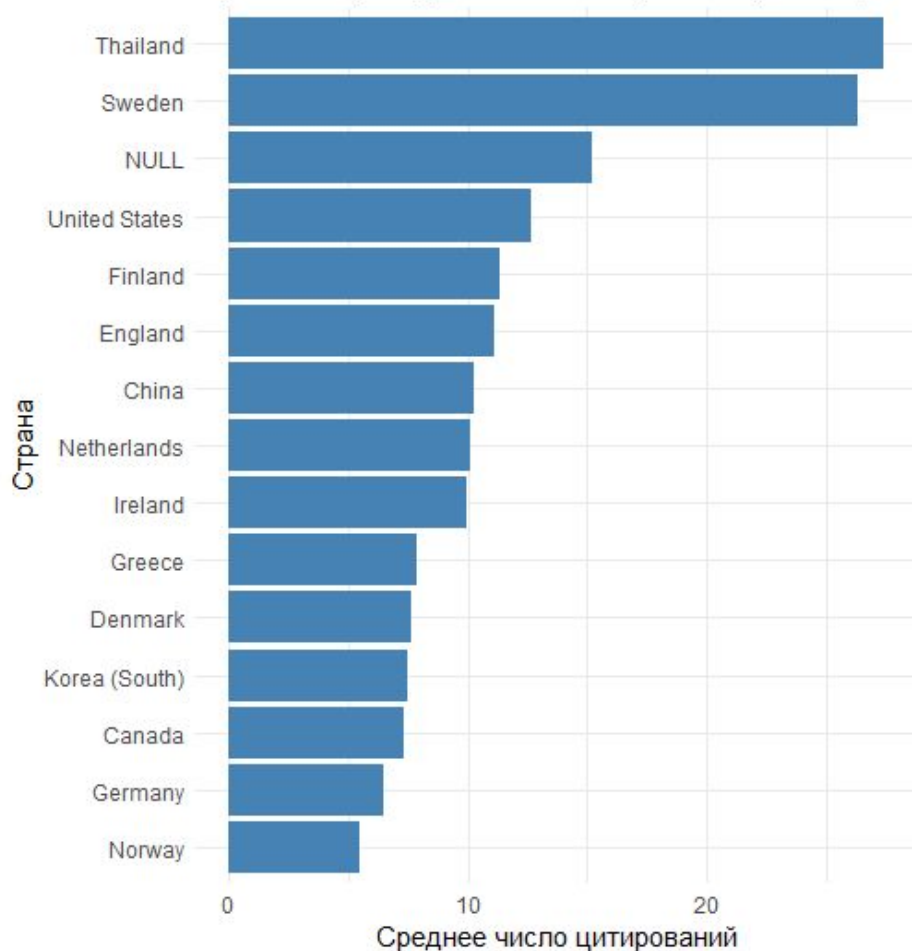


**График распределения количества авторов к количеству публикаций.** В датасете есть статьи с максимальным количеством авторов 1065, это либо выброс, либо большие коллаборации. Было решено для графика сократить до 75 авторов максимум.

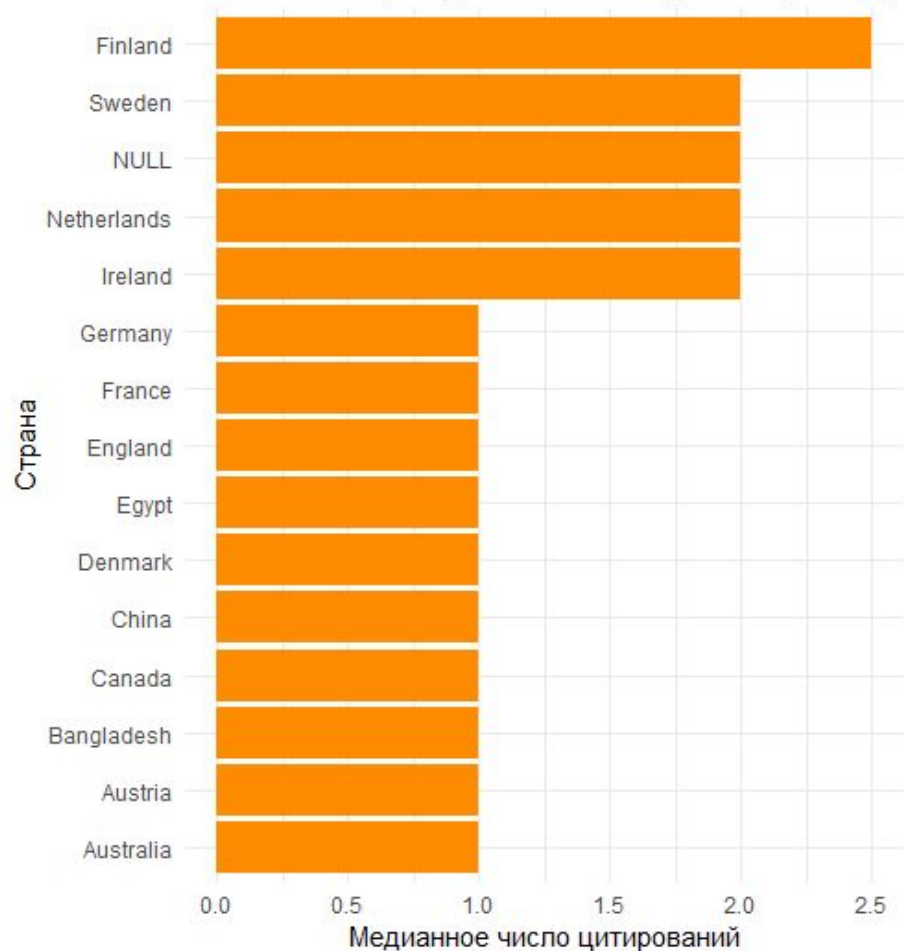
Распределение количества авторов



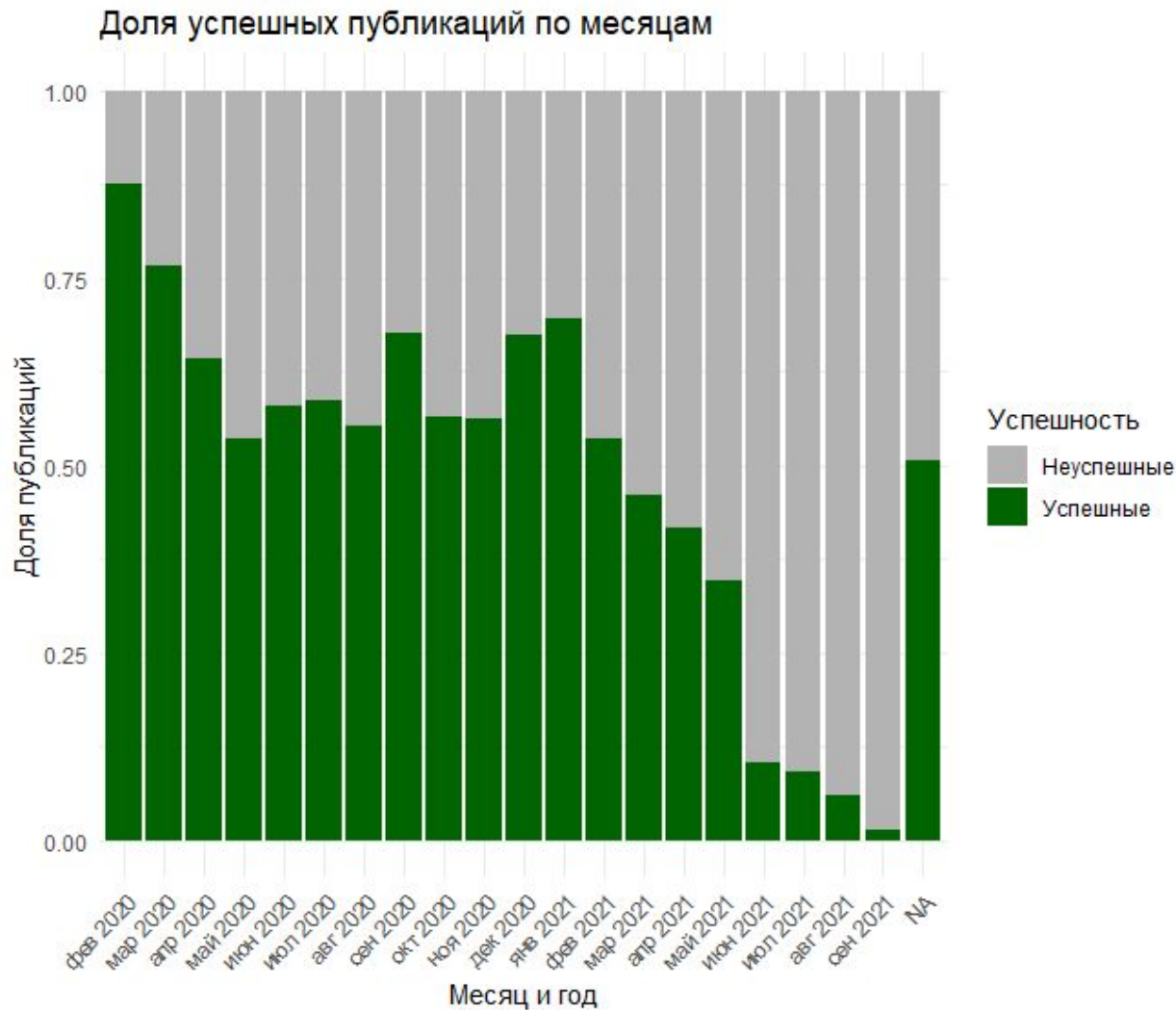
Средняя цитируемость по странам (топ-15)



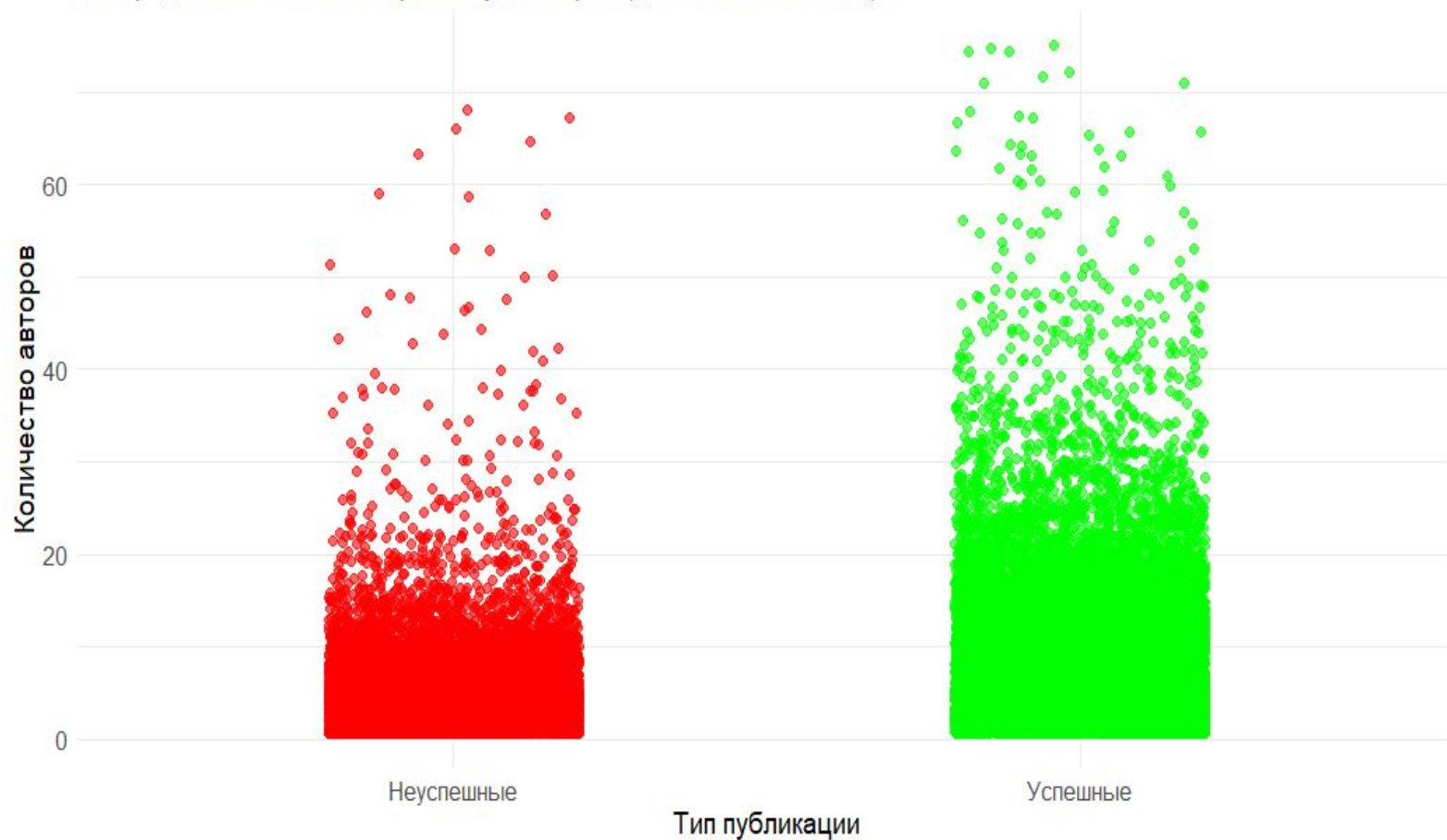
Медианная цитируемость по странам (топ-15)



- Доля успешных публикаций по месяцам (большой процент в феврале, марте 2020 и январе 2021).
- Однако, если соотнести с графиком количества статей по месяцам, то видно что наибольшая публикативная активность из этих месяцев была в декабре 2020 и январе 2021. Поэтому далее использовались данные за эти месяцы.

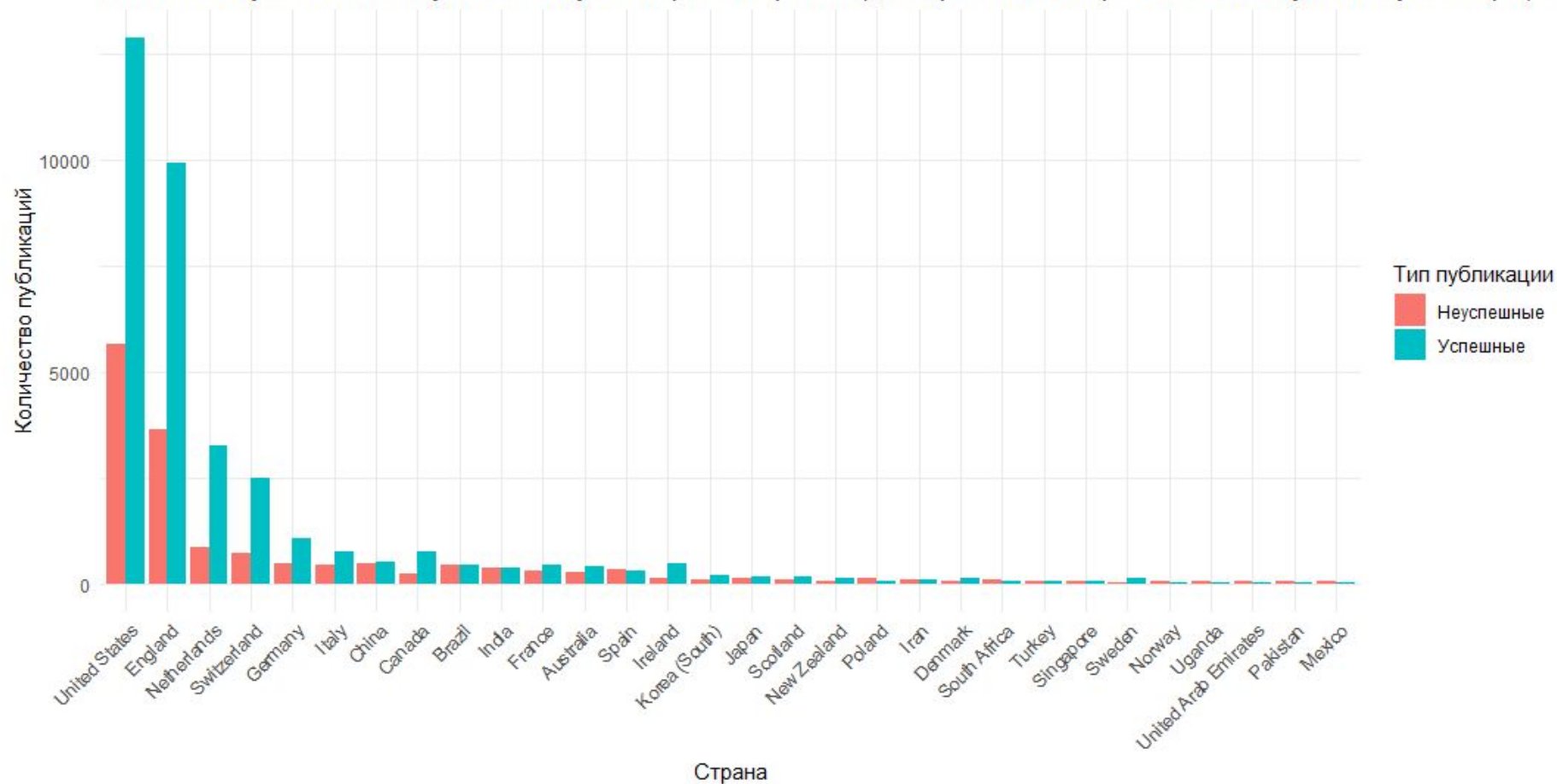


Распределение числа авторов в публикациях (дек 2020 и янв 2021)



Уже заметна небольшая связь числа авторов и успешности публикации

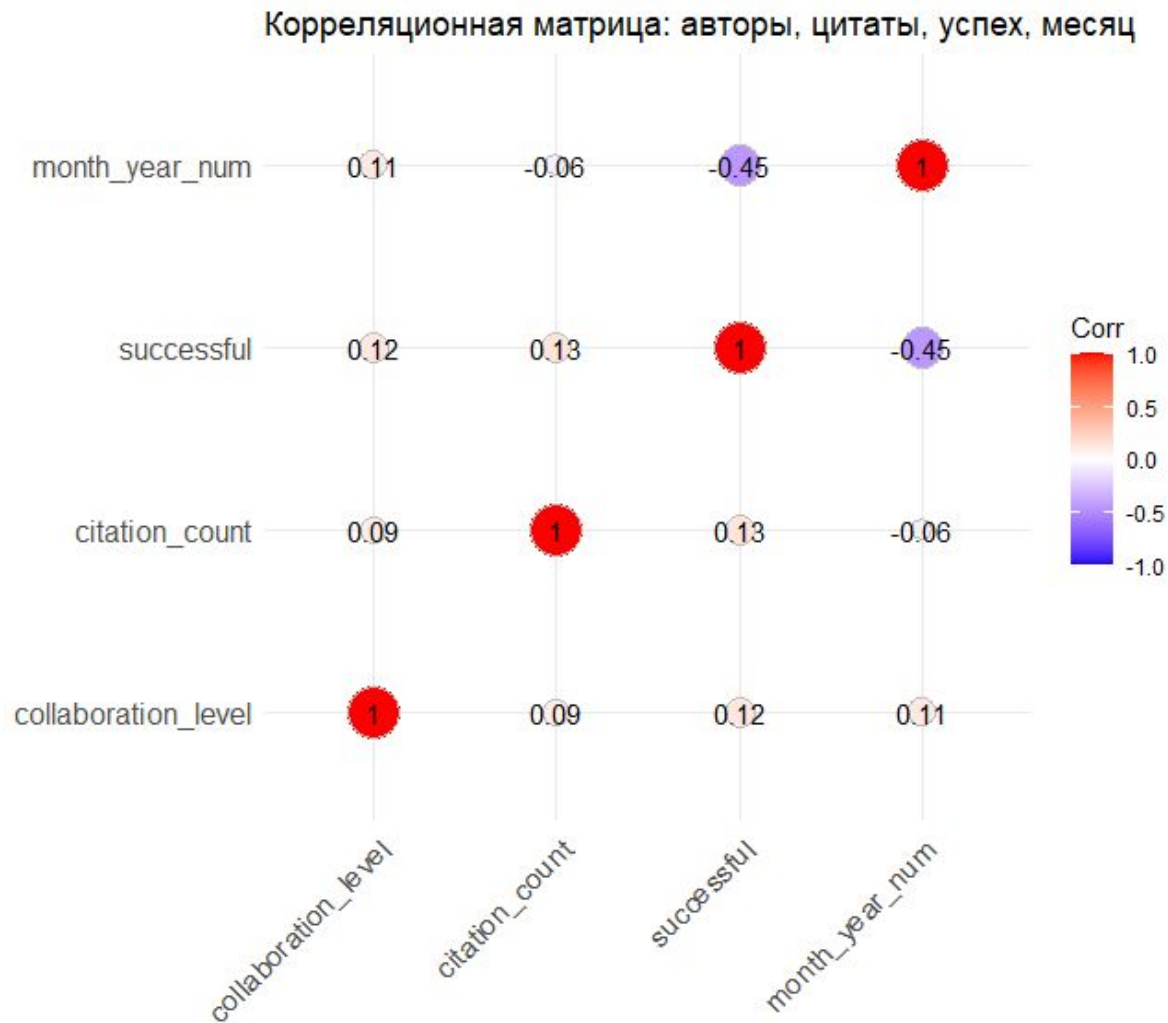
Количество успешных и неуспешных публикаций по странам (декабрь 2020/январь 2021, минимум 100 публикаций)



Самые успешные публикации были из США и Великобритании.

На графике показаны парные корреляции между ключевыми количественными переменными:

- **collaboration\_level** — количество авторов статьи
- **citation\_count** — число цитирований
- **successful** — бинарная переменная успешности (выше медианы по цитируемости)
- **month\_year\_num** — порядок месяца публикации



# Выводы

**Время публикации имеет наибольшее значение — статьи, опубликованные в начале пандемии, стали более успешными. Это подтверждает отрицательная корреляция даты с успешностью.**

**Командная работа помогает быть успешнее: статьи с большим числом авторов немного более успешны.**

**Портрет успешного автора — опубликовался в начале пандемии, работает в команде, из США или Великобритании**

