МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение

высшего образования

«КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Департамент развития дополнительного образования

# ДПП ПП «Основы анализа данных в научной деятельности»

**ОТЧЕТ**

по практике

|  |
| --- |
| Слушатель Рамазанова Миляуша Илдаровна  (ФИО) |
| Слушатель Ефимова Диляра Маратовна  (ФИО)  Слушатель Нурягдыева Эмине  (ФИО)  Слушатель Гайнутдинова Аделина Рустемовна  (ФИО)  Слушатель Шевченко Роман Васильевич  (ФИО)  Слушатель Минязева Ирина Салаватовна  (ФИО) |
| Руководитель практики от ООО «Нетология»:  Перепелкина Наталья Александровна,  руководитель продюсерского центра департамента по работе с органами государственной власти ООО «НЕТОЛОГИЯ»    Дата сдачи отчета 30.05.2025 |
|  |
|  |

Казань, 2025 год

СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 3](#_Toc197358339)

[ЦЕЛИ И ЗАДАЧИ ПРОЕКТА 5](#_Toc197358340)

[ТЕХНОЛОГИЧЕСКИЙ СТЕК 7](#_Toc197358341)

[СПИСОК ЛИТЕРАТУРЫ И РЕСУРСОВ 8](#_Toc197358342)

# ВВЕДЕНИЕ

Практика проходила с 28 апреля по 30 мая 2025 года в компании ООО «Нетология». Это российская компания и [образовательная онлайн-платформа](https://ru.wikipedia.org/wiki/%D0%94%D0%B8%D1%81%D1%82%D0%B0%D0%BD%D1%86%D0%B8%D0%BE%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5" \o "Дистанционное обучение), созданная в 2011 году. Одна из ведущих российских компаний онлайн-образования, участник проекта создания и обеспечения функционирования инновационного центра «Сколково» (выписка из Реестра участников №В-4795 от 03.06.2024).

Образовательная деятельность «Нетологии» ведётся на основании [государственной лицензии №Л035-01298-77/00180129](https://islod.obrnadzor.gov.ru/rlic/details/997542a2-0e2b-956f-9413-7ec2f35f38d8/" \t "_blank) от 14 апреля 2021 г.

«Нетология» является аккредитованной ИТ-компанией на основании решения Министерства цифрового развития, связи и массовых коммуникаций Российской Федерации о предоставлении государственной аккредитации организации, осуществляющей деятельность в области информационных технологий от 29.03.2022 № АО-20220328-3935807532-3.

Целью прохождения практики является отработка навыков применения языка R для решения профессиональных задач.

В качестве темы проекта было решено провести анализ пакета данных с обзорами научных исследований на тему "Новая короновирусная инфекция (COVID-19) за период 2020-2021 г.".

Актуальность темы проекта:

Пандемия COVID-19 вызвала резкий всплеск научных публикаций. В условиях информационного взрыва появилась необходимость быстро оценивать качество и значимость новых исследований. Понимание факторов, влияющих на "успех" публикации (например, число цитирований), помогает исследователям, редакторам и аналитикам выявлять важные научные тренды и принимать обоснованные решения.

С помощью данного проекта мы попытались решить следующую проблему:

Существуют тысячи научных статей по теме COVID-19, но лишь часть из них получает широкое признание в научном сообществе. Возникает вопрос: какие характеристики публикации (например, число авторов, дата публикации, кооперация между странами и др.) могут влиять на её успех?

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

# ЦЕЛИ И ЗАДАЧИ ПРОЕКТА

Цель проекта:

Провести исследовательский анализ данных (EDA) по публикациям CORD-19 и выявить закономерности, связанные с успешностью и цитируемостью научных статей.

Задачи:

* 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.

Ход реализации проекта:

1. Загрузка данных:

Импортирован датасет CORD-19 с Kaggle: метаинформация о научных публикациях по COVID-19.

2. Понимание данных:

Исследованы ключевые поля: авторы, дата публикации, количество цитирований, журналы и др.

3. Очистка данных:

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. Визуализация данных:

Построены гистограммы, боксплоты, плотности распределений, временные тренды и корреляционные матрицы.

6. Анализ основных статистик:

Проанализированы распределения цитирований, уровни кооперации и временная динамика публикаций.

7. Исследование зависимостей:

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

8. Поиск аномалий и выбросов:

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

9. Итеративный анализ:

Проводилось уточнение переменных, пересчёт успешности и пересмотр порогов.

10. Документация и представление данных:

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

# ТЕХНОЛОГИЧЕСКИЙ СТЕК

В работе над проектом использовался язык программирования R.

Код проекта размещён по ссылке: https://github.com/dilyarachan0930/ProjectCOVID19-study/blob/77096b6b263d78b9d06de2a95ff7fefad5e33f36/script2.R

# СПИСОК ЛИТЕРАТУРЫ И РЕСУРСОВ

**Основная литература:**

1. Кабаков, Р. R в действии. Анализ и визуализация данных в программе R / Р. Кабаков. - М.: ДМК, 2016. - 588 c.
2. Люк Д. Анализ сетей (графов) в среде R. Руководство пользователя. М.: ДМК Пресс, 2016.
3. Мастицкий, С.Э. Статистический анализ и визуализация данных с помощью R / С.Э. Мастицкий. - М.: ДМК, 2015. - 496 c.
4. Мэтлофф, Н. Искусство программирования на R / Н. Мэтлофф. – СПб.: Питер, 2019. – 448 с.
5. Шипунов А.Б. и др. Наглядная статистика. Используем R! М.: ДМК Пресс, 2014. 298 с.

**Перечень Интернет-ресурсов, необходимых для освоения дисциплины:**

1. https://cran.r-project.org/ Официальный сайт языка R
2. https://www.rstudio.com/resources/learn/ Обучающие ресурсы от разработчиков RStudio
3. https://rstudio.cloud/ Онлайн-среда для работы с R
4. https://rmarkdown.rstudio.com/ Документация и примеры по R Markdown
5. https://ggplot2.tidyverse.org/ Документация по библиотеке ggplot2 для визуализации данных
6. https://github.com/cran/ Репозиторий исходного кода пакетов CRAN на GitHub