Всероссийский конкурс научно-технологических проектов "Большие вызовы"

**Web-ресурс обработки, визуализации и предсказания статистики заболеваемости и смертности от covid-19.**

Ульяновск 2022

**Содержание:**

1. **Введение.**
2. **Сбор и анализ данных.**
3. **Преобразования данных.**
4. **Визуализация данных.**
5. **Нормирование.**
6. **Корреляция.**
7. **Линейная регрессия и первые прогнозы.**
8. **Влияние и статистика вакцинации.**
9. **Итоги.**
10. **Перспективы.**

**Введение:**

В настоящее время весь мир терроризирует коронавирус, из-за которого уже заболели сотни миллионов человек, из которых миллионы - это летальные исходы. Какое-то время назад многие страны мира боролись с этим заболеванием на пределе возможностей: постоянно забитые больницы, измученные врачи, работающие по две смены, и, конечно же, постоянная разработка вакцины против covid-19. Всё это генерирует и позволяет собирать большое количество данных об этом лёгочном заболевании, которыми мы и воспользуемся в нашей исследовательской работе для визуализации и обработки. А сделаем мы это для более удобной и наглядной оценки.

Как было уже сказано выше, данные, которые мы используем в работе постоянно актуализируются и обновляются из открытых источников, а именно из популярного ресурса [**Our World in Data**](https://ourworldindata.org/). Это и позволит нам сделать прогноз на несколько дней вперёд.

Исходя из всего этого, **гипотеза**, которую мы хотим проверитьтакова: «Количество вакцинировавшихся напрямую коррелирует с количеством заболевших, а страны с населением, либо правительством с более мягким отношением к covid-19 имеют более высокие показатели процентальной заболеваемости, относительно населения».

**Проблематика:**

Прогнозирование - это очень важная часть моделирования нынешней ситуации в любой стране. Если смочь правильно спрогнозировать будущие препятствия, то возможно к ним подготовиться заранее. А также, здоровье - это самый важный человеческий ресурс, который очень дорого и трудозатратно восстанавливать. Никто бы не хотел потратить его впустую, для этого все мы должны понимать уровень опасности болезни, которая нас постигла, а также все мы должны уметь ей противостоять.

**Цели и задачи:**

Целью нашего исследования является визуализация смертности и заболеваемости в разных странах, а также поиск и сравнение стран с похожими ситуациями. На основе этих данных, прогноз количества заболевших на несколько дней вперёд и выяснение, действительно ли прививка влияет на количество заболевших.

**Основные понятия:**

**Dash** - это библиотека Python, которая позволяет визуализировать данные, а также выводить их в отдельном окне, с которым можно взаимодействовать.

**Plotly** - это библиотека Python, позволяющая визуализировать табличные данные для удобоваримости.

**Numpy** - это библиотека Python, которую мы использовали для более удобного хранения и сортировки данных.

**Pandas** - это библиотека Python для упорядочивания и редактирования исходных данных.

**Sklearn** - это набор библиотек, позволяющий создать простенький искусственный интеллект, который сможет сделать предсказание.

**Корреляция** - это линейная зависимость между двумя объектами, определяющаяся диапазоном от -1 до 1.

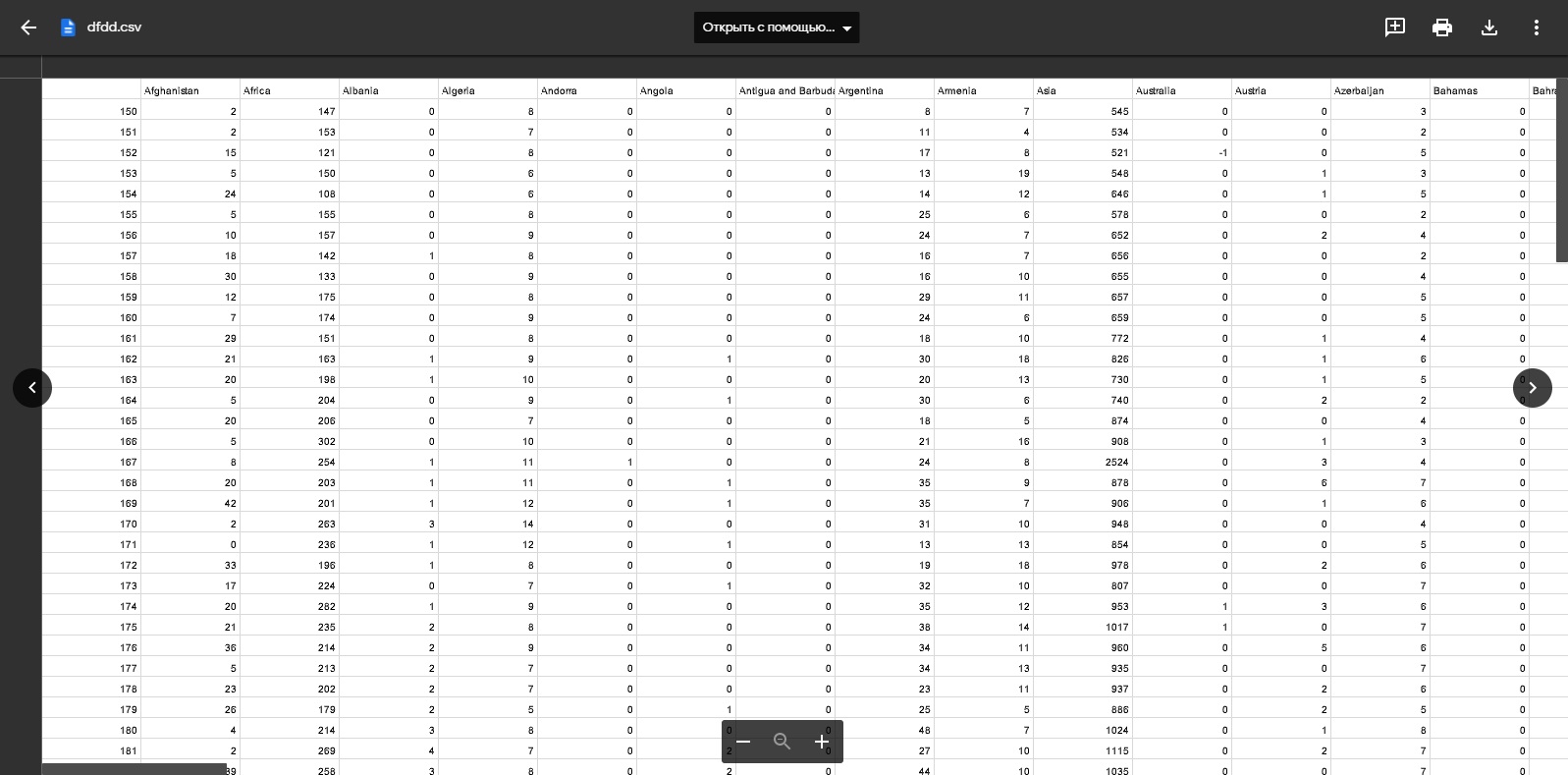
**Нормирование** - это усреднение всех данных с целью более удобного сравнения.

**Сбор и анализ данных:**

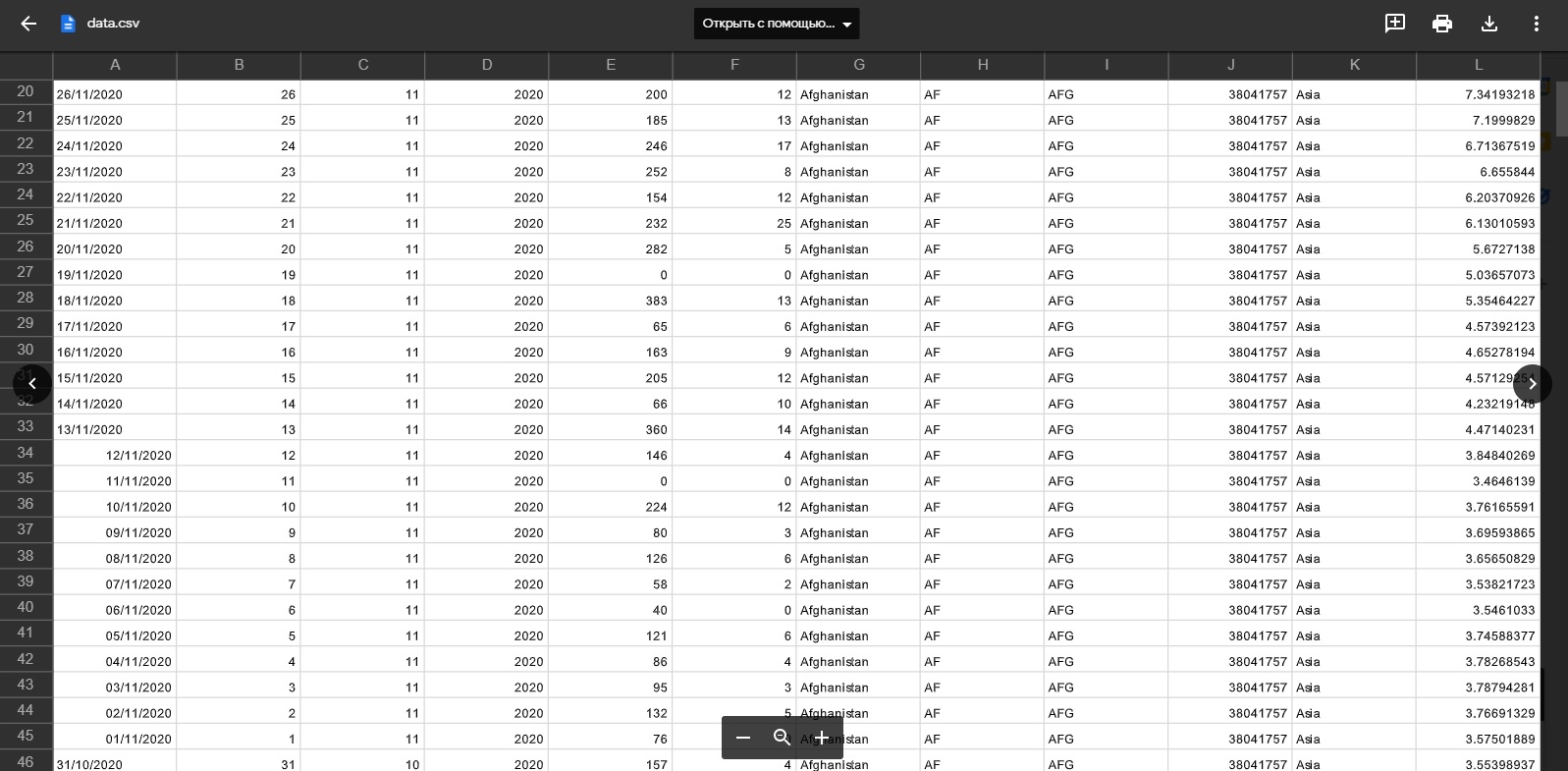
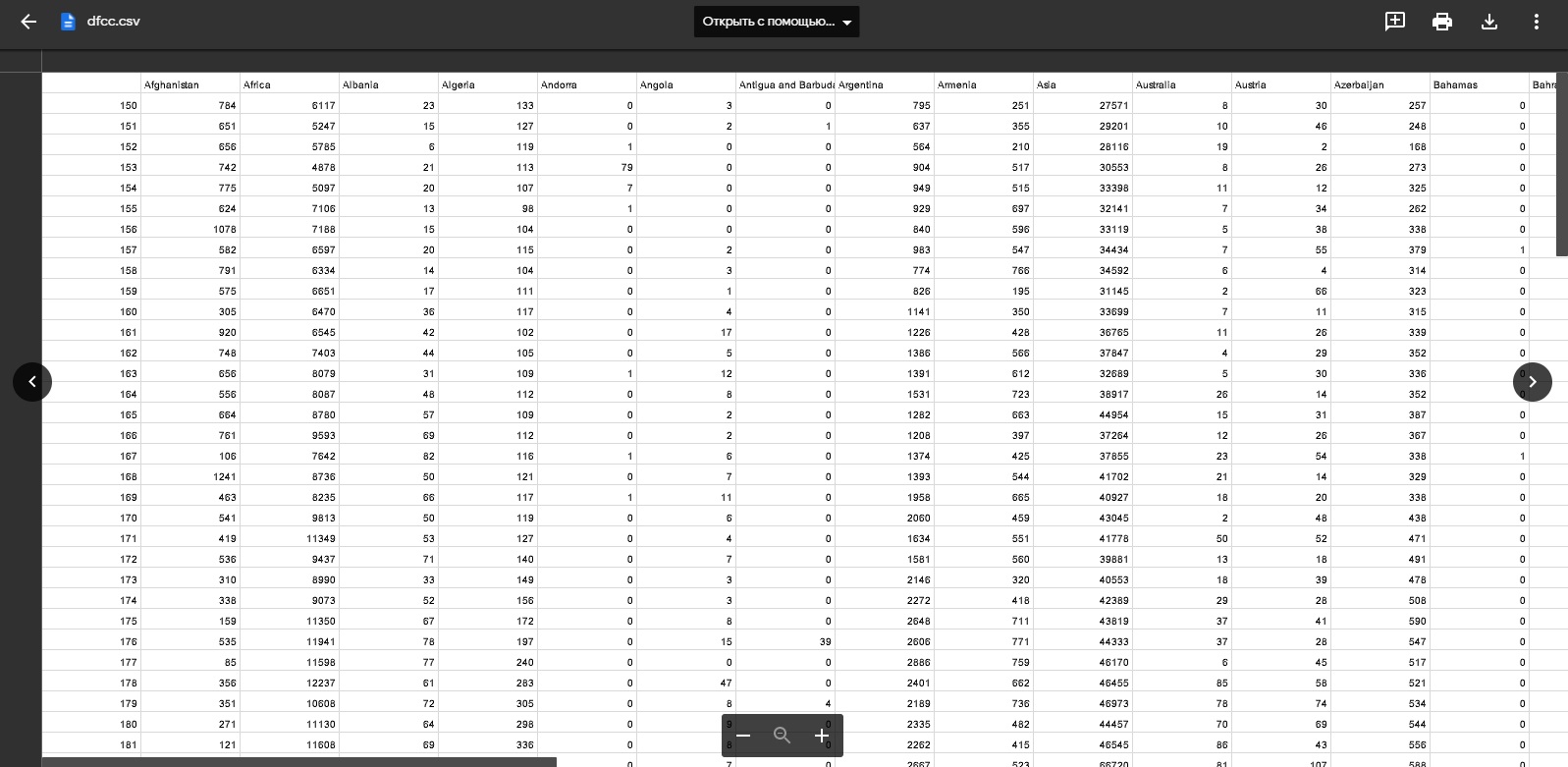
Начали свой поиск мы с небольшого файла, в котором находились данные о заболеваемости и смертности на 2020-ый год. Для начала этого было достаточно, но лишь для начала. Со временем, мы поняли, что актуальность этих данных очень низкая, так что, перешли к следующему файлу, в котором уже были данные по сей день. Это очень помогла в работе с данными, так как мы стали лучше представлять то, о чём идёт речь и данные стали более осязаемыми, даже новости пошли в бой.

Но это был один из самых простых наших шагов. Дальше же пошла настоящая работа.

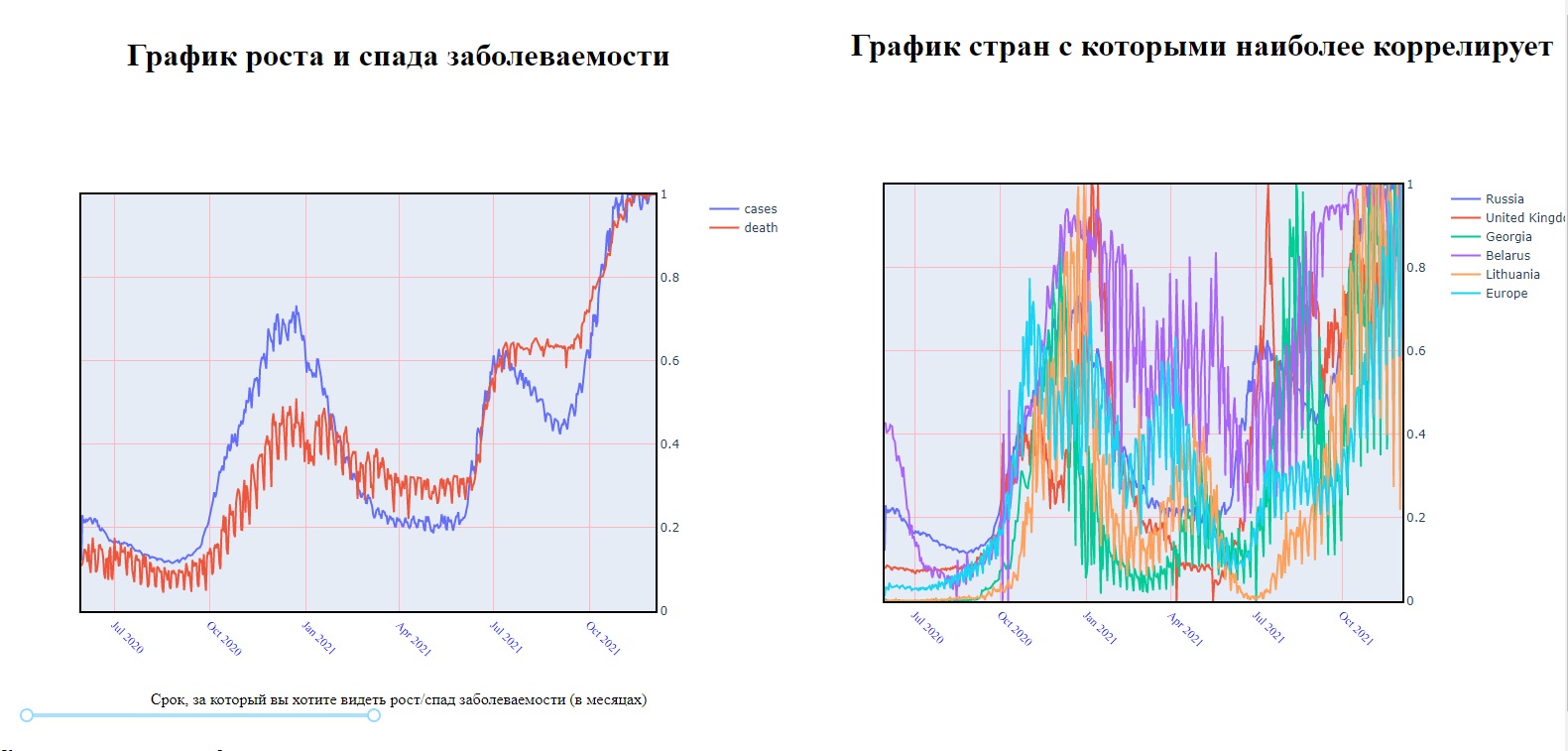
**Обработка данных:**

Для преобразования всех наших данных мы использовали Python и конкретно его библиотеку - Pandas. Так как изначально у нас был файл на 61901 строчку и 12 столбиков со статистикой разных стран,

с которым было абсолютно неудобно работать, мы стали его преобразовывать в две разные таблицы - это таблица смертности и таблица заболеваемости. Которые у нас вышли уже на 554 строчки и 181 колонку

 Кроме того, эти две таблицы у нас были отсортированы по дням, что нам очень помогло в будущем. Также, несколько стран, которые не имели веса в дальнейшем были пропущены, так как просто нагружали наш код лишними операциями. Все эти действия нам позволили более продуктивно работать, так что, определённо, это стоило потраченного времени.

**Визуализация данных:**

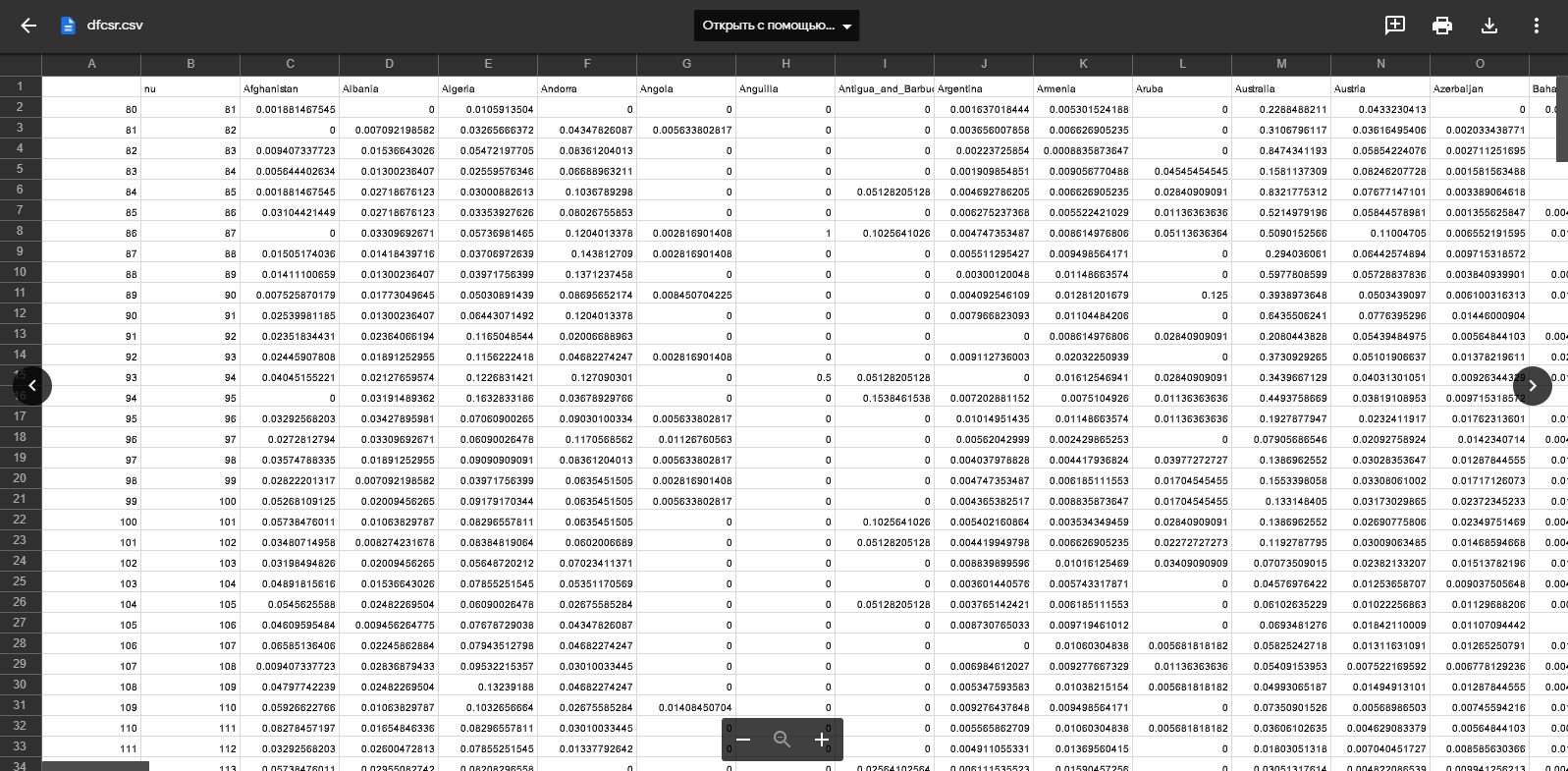
Для визуализации данных изначально мы использовали библиотеку Pandas, но она не позволила бы нам вывести все наши графики на отдельную интернет страничку, чтобы мы могли более наглядно показать все данные, которые были получены и обработаны, из-за чего, в итоге мы перешли к набору библиотек Dash и Plotly, что уже позволило задуманное нами воплотить в реальность. 

И тут, мы начали сталкиваться с одной проблемой за другой.

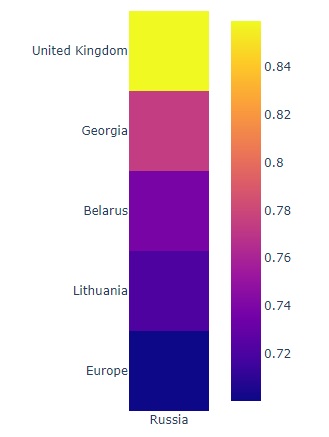
Так как мы хотели сделать сайт, он должен быть хоть немного, но интерактивным. Для этого мы сделали возможность выбора стран в специальной панели. И выбор срока, за который можно видеть количество заболевших. Для работы этого всего нужна функция обратного вызова, которая и была основным источником нашей головной боли.

**Нормирование:**

Со временем нам пришлось пронормировать данные, чтобы мы могли сравнить графики заболеваемости разных стран и графики заболеваемости одной страны с графиком смертности всё той же страны, а так как без нормирования это было невозможно, мы пошли искать варианты, которые нам бы помогли с этим. Так как мы поняли, что вариантов было куча, мы попробовали не сильно сложный вариант - это было линейное нормирование(x - x(min) / x(max) - x(min)), оно нам показалось подходящим, но мы решили попробовать ещё один вариант - логорифмизация данных(log10(x)), этот вид нормирования нам уже показалось менее подходящим, так что, остановились мы на первом варианте. Из-за чего все наши данные находились в диапазоне от 0 до 1, что позволило их сравнивать на графиках.

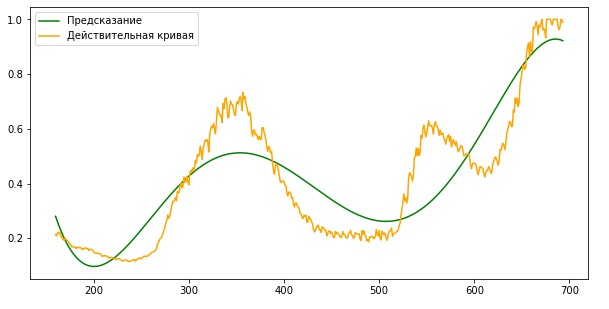


**Корреляция:**

После нормирования и визуализации нормированных данных, мы пришли к выводу о том, что на глаз нам никак не определить схожие страны, лишь с большой погрешностью, а для того, чтобы не заниматься самодурством, мы пришли к выводу, что нам нужно знать корреляцию между странами. Но всё также, и тут без трудностей не обошлось. И так мы нашли два вида корреляции, которые нам подходили - это с помощью Numpy и с помощью Pandas. Конечно же, просто так нам ничего не давалось, ведь нам надо было узнать как-то самые похожие страны, но и Numpy, и Pandas нам предлагали лишь корреляцию между всеми странами. Но, это было не сложной задачей, так как мы уже были уверенными пользователями на тот момент, нам лишь пришлось отсортировать полученные данные и взять первые 5 коррелирующих страны с той, которую мы прокоррелировали. И та-дам, у нас есть самые похожие страны.

**Обучение линейной регрессии:**

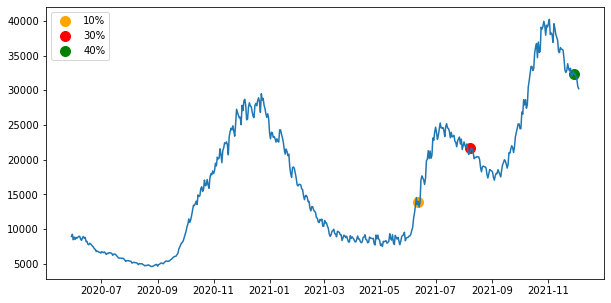
Итак, одной из наших изначальных задач было предсказание на несколько дней вперёд, теперь мы можем этим заняться. У нас есть самые коррелирующие страны, они и будут нашей основой для обучения линейной регрессии, но не все, конечно же, пяти будет достаточно. Что нам требуется дальше? Нам понадобится набор библиотек sklearn, с помощью которых мы сможем написать большой код простейшего искусственного интеллекта, который сможет построить примерный график, используя данные, на которых он учится, а это означает, что его можно как-то описать, что в свою очередь означает, что его можно построить дальше. Это и будет нашим предсказанием, но с небольшой погрешностью.

****

**Статистистические данные о вакцинации:**

Так как у нас уже была вся информация, мы решили подвести некоторые интересные статистические данные.

Мы решили узнать, в какой момент в России вакцинировалось 10%, в какой 30% и в какой 40% человек. Узнать мы это смогли, благодаря тому же открытому ресурсу [Our World In Data](https://ourworldindata.org/), благодаря которому мы узнали и о заболеваемости.

А также, мы можем узнать сезонность и частоту заболеваемости по дням недели и месяцам, что мы и сделали. И вот, что мы получили: 