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 вирусном заболевании, которыми мы и воспользуемся в нашей исследовательской работе для визуализации и обработки. А сделаем мы это для более удобного и наглядного сравнения заболеваемости и смертности в разных странах.

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/). Что позволит нам пользоваться самой свежей информацией и создавать прогнозы на несколько дней вперёд с помощью регрессии.

Исходя из всего этого, **гипотеза**, которую мы хотим проверитьтакова: «Имея на руках лишь данные, которые дают нам информацию про заболеваемость и смертность граждан в той или иной стране, можно вычислить корреляцию между разными странами, с помощью неё найти несколько самых похожих стран и воспользоваться ими для обучения линейной регрессии и, с помощью регрессии, предсказывать заболеваемость в конкретной стране на несколько дней вперёд».

**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, поддерживающий множество методов, в том числе и регрессию.

**Регрессия** - это зависимость, устанавливающая соответствие между случайными переменными, то есть математическое выражение, отражающее связь между зависимой переменной *у* и независимыми переменными *х*.

**Полином** - это то же самое, что и многочлен, либо же какая-то сумма одночленов.

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

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

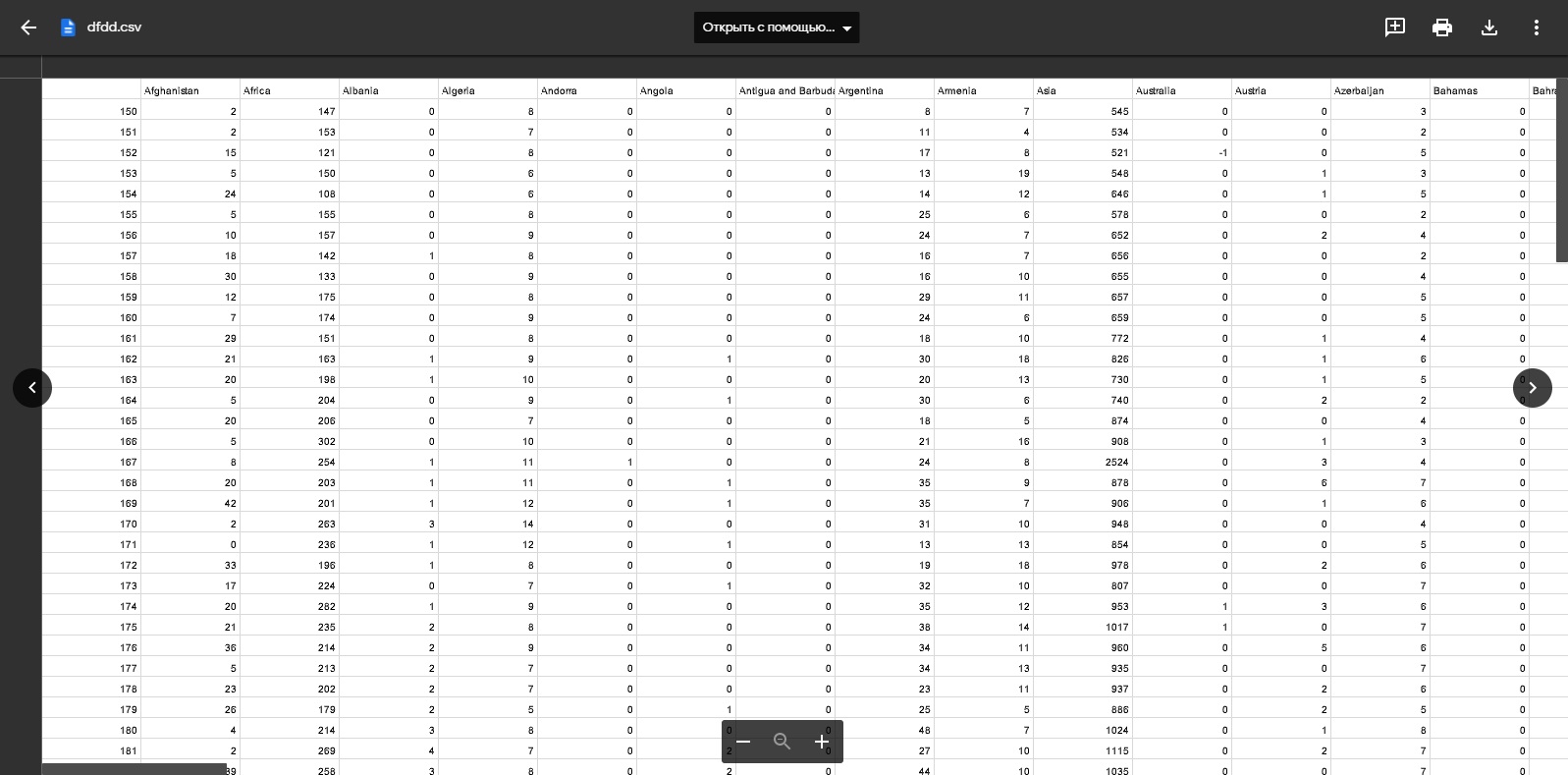
**Сбор и анализ данных:**

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. Этот файл очень помог нам 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.

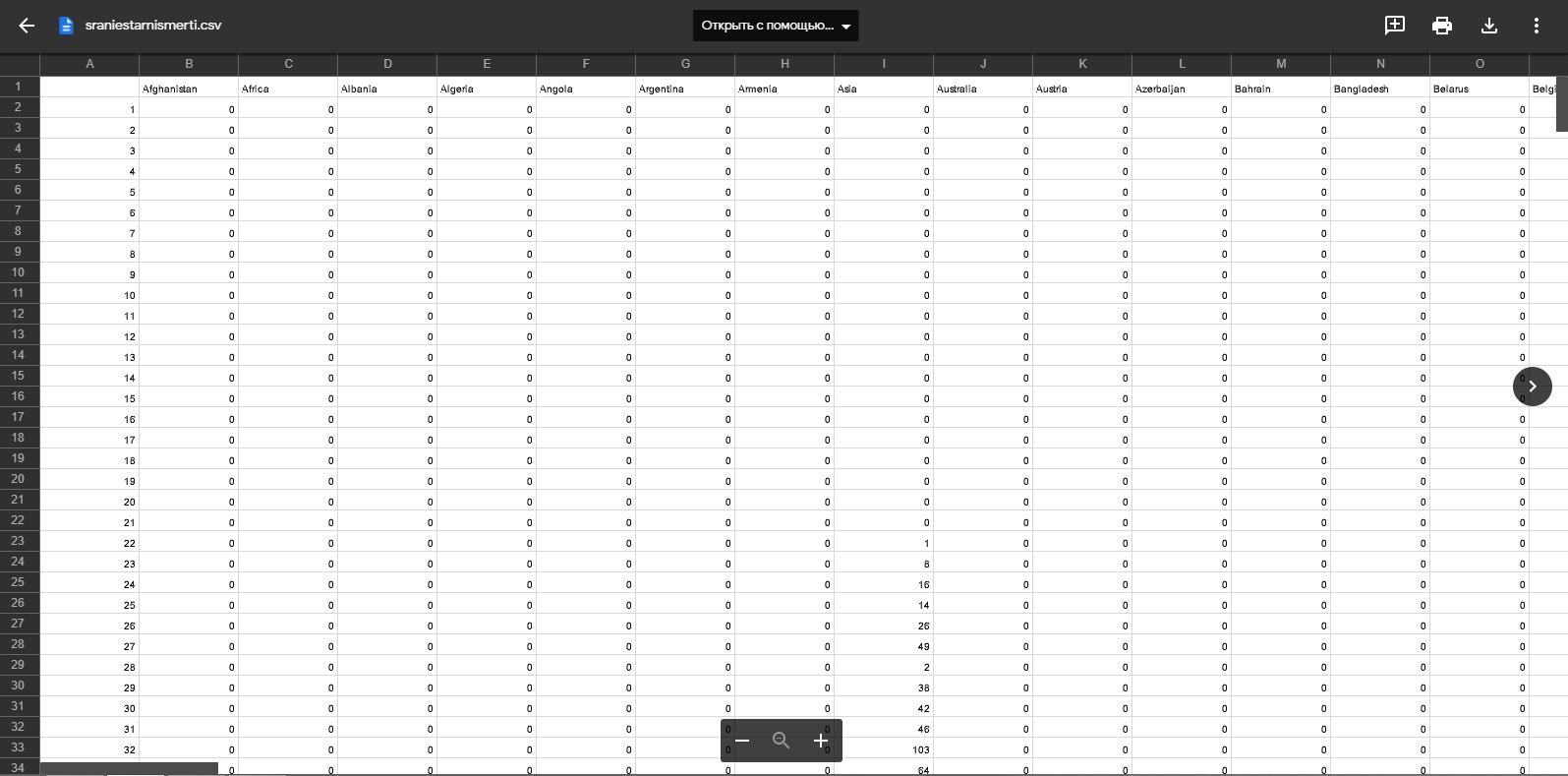
**Обработка данных:**

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 (**Рисунок 1**) со статистикой разных стран, с которым было абсолютно неудобно работать, мы стали его преобразовывать в две разные таблицы - это таблица смертности (**Рисунок 2**) и таблица заболеваемости (**Рисунок 3**). Которые у нас вышли уже на 706 строчек, которые описывали каждый день и 130 колонок, которые описывали страны, а в каждой ячейке у нас были смерти за конкретный день в конкретной стране, например, так выглядела часть таблицы заболеваемости:

|  | Russia | World |
| --- | --- | --- |
| 401 | 16379.0 | 381773.0 |
| 402 | 15808.0 | 405792.0 |

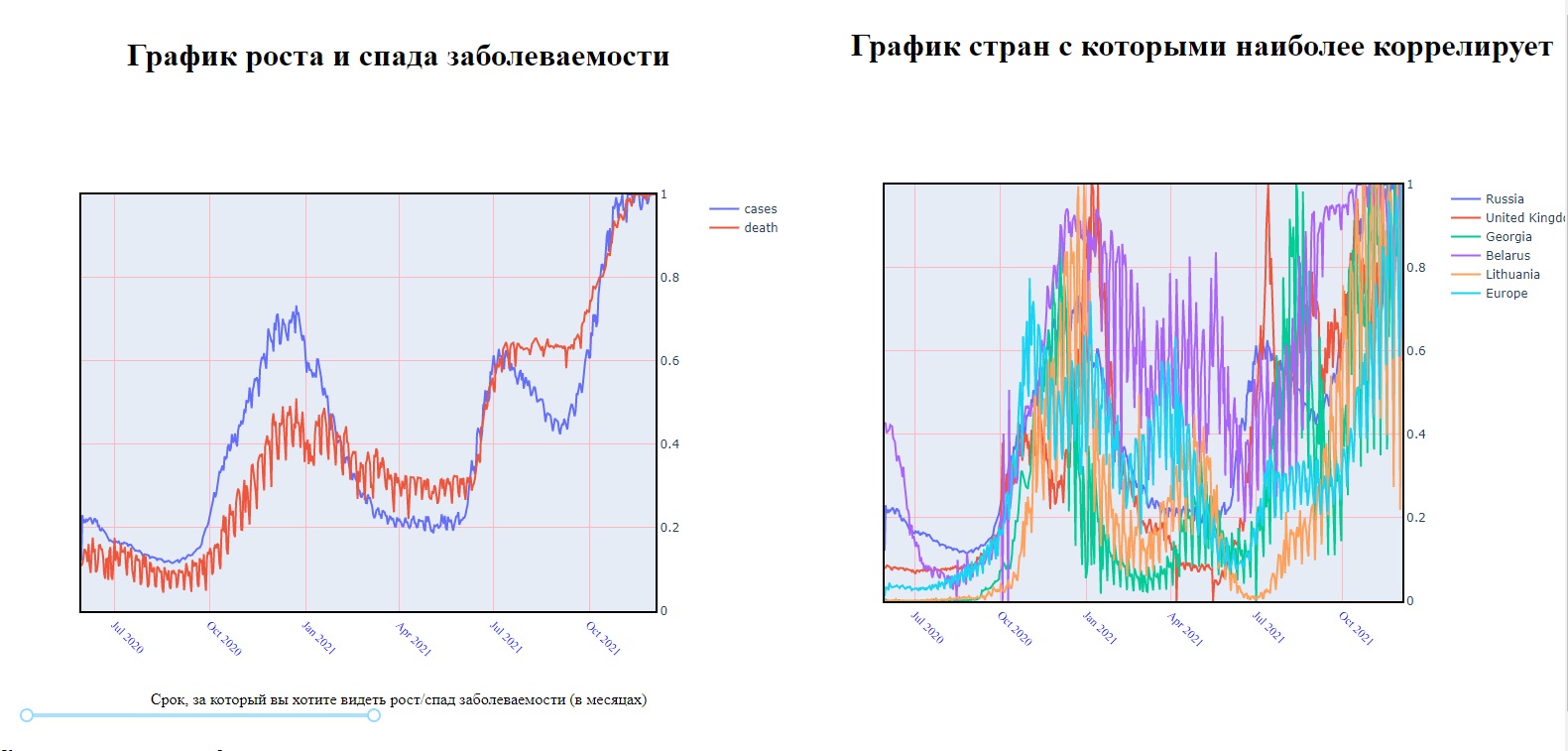


**Рисунок 1.**

**Рисунок 2. Рисунок 3.**

Кроме того, эти две таблицы у нас были отсортированы по дням, что нам очень помогло в будущем. Также, несколько стран, которые не имели в себе данных, а если и имели, то были заполнены нулями, 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. А выглядело это вот так:

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:**

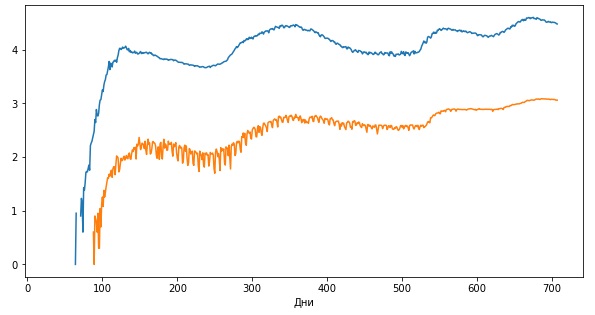
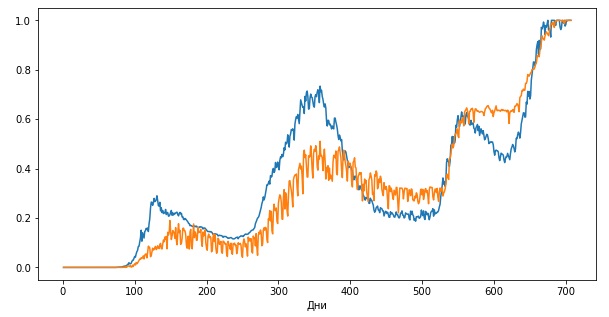
Со временем нам пришлось пронормировать данные, чтобы мы могли сравнить графики заболеваемости разных стран и графики заболеваемости одной страны с графиком смертности всё той же страны, а так как без нормирования это было невозможно, мы пошли искать варианты, которые нам бы помогли с этим. Так как мы поняли, что вариантов было куча, мы попробовали не сильно сложный вариант - это было линейное нормирование (**Рисунок 4**), оно нам показалось подходящим, но мы решили попробовать ещё один вариант - логорифмизация данных (**Рисунок 5**), этот вид нормирования нам уже показался менее подходящим, так как не позволял визуально данные, а этого мы и добивались, так что, остановились мы на первом варианте.

В результате применения линейного нормирования все наши данные находились в диапазоне от 0 до 1, что позволило их сравнивать на графиках. После применения данные в нашей таблице изменились (**Рисунок 6**) и стали выглядеть так:

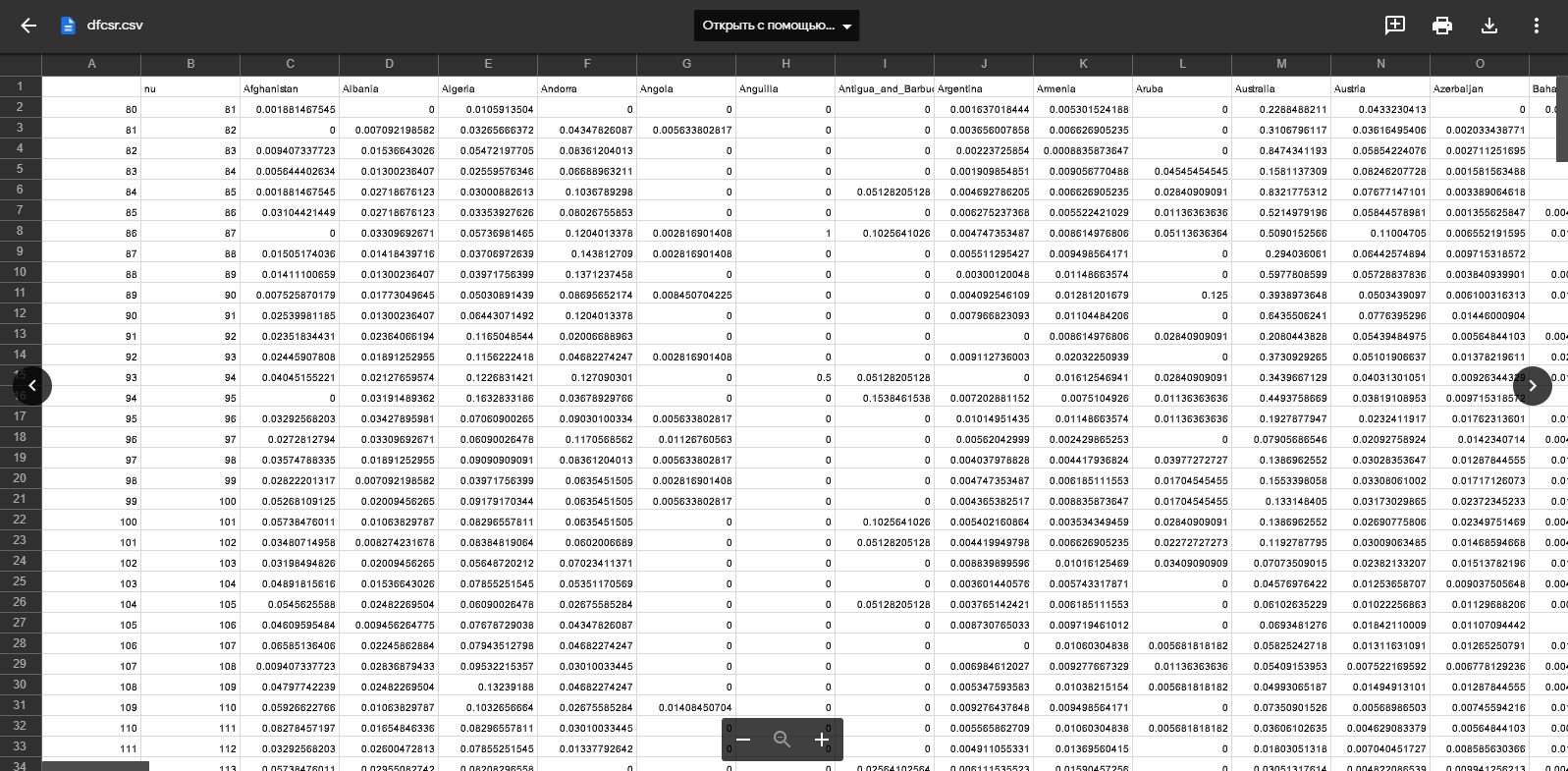
|  | Russia | World |
| --- | --- | --- |
| 401 | 0.407336483461825 | 0.420321065211623 |
| 402 | 0.393136035811987 | 0.446765291663776 |

**Линейное нормирование:** , где **X** - это данные о заболеваемости и смертность в какой-то день.

**Логорифмизация:** , где **X** - это данные о заболеваемости и смертность в какой-то день.

****

**Рисунок 4. Рисунок 5.**



**Рисунок 6.**

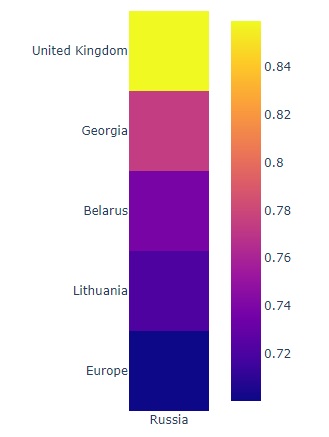
**Корреляция:**

Для чего нам нужна была корреляция? Так как мы решили спрогнозировать количество заболевших на несколько дней в будущее, то нам необходимо было узнать несколько стран, которые бы были схожи по процентальной заболеваемости, с той страной, количество заболевших в которой мы бы хотели узнать. А для того, чтобы найти процентальные значения для каждой страны, нам бы потребовалось заново рассчитывать количество заболевших и сравнивать с количеством населения в стране, а этих данных у нас не было. Так что, было принято решение взять нормированные данные, которые отлично подходили, так как были в одном и том же диапазоне и сравнимы друг с другом.

Мы попробовали воспользоваться корреляцией, предлагаемой нам библиотекой Numpy, но это было безуспешно, так как наши данные не подходили под требования Numpy, чтобы с ними работать, но также прокоррелировать данные могла и библиотека Pandas, причём успешно, так как данные хранились и преобразовывались именно при помощи этой библиотеки.

После проделанной работы у нас была таблица со значениями коррелируемых стран для каждой страны, которая выглядела как-то так:

|  | Algeria | Angola |
| --- | --- | --- |
| Algeria | 1.000000 | 0.162005 |
| Angola | 0.162005 | 1.000000 |



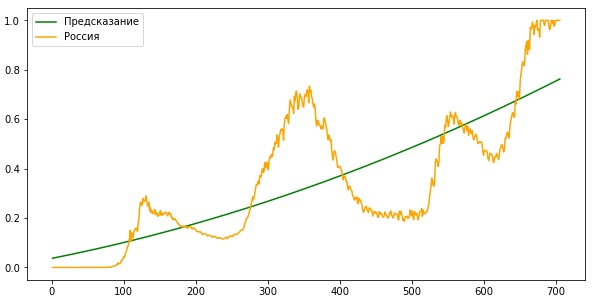
Дальше, мы искали страны, которые коррелируют со страной, которую мы задали лучше всего и строили по ним тепловую карту, которую можно увидеть слева, чтобы это выглядело наглядно и понятно.

**Обучение регрессии:**

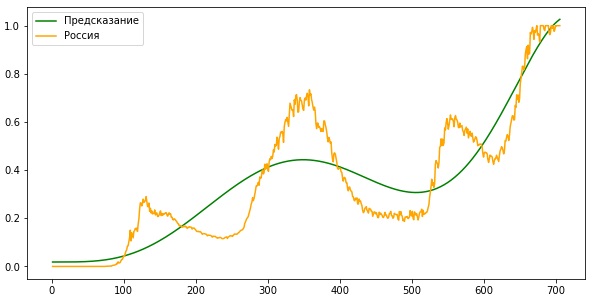
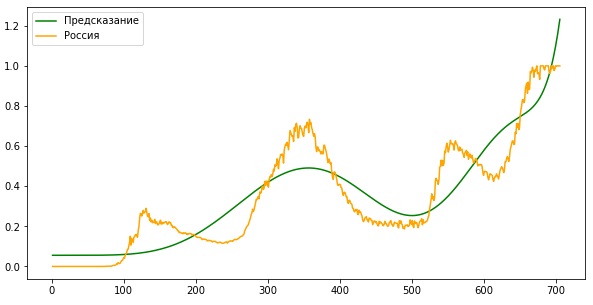
Последний вопрос, который так и не остался раскрытым - это вопрос, связанный с предсказанием? Возможно ли это, а если и возможно, то насколько точно?

Так как все данные были подготовлены, мы приступили к созданию полиномиальной регрессии с помощью набора библиотек sklearn, которые в себе поддерживали множество методов, в том числе и регрессию. Для начала, нам надо написать небольшой кусок кода, который будет обучаться по имеющимся данным и строить свой полином, максимально приближенный к ним.

Далее, нам требуется выбрать подходящую степень полинома, так как есть более удачные версии полинома, которые проходят максимально близко к нашему графику, а есть и менее удачные. Выбирать мы будем с помощью RMSE или же средне-квадратичной ошибке, которая позволит нам найти наименьшее отклонение, например, полином второй степени (**Рисунок 7**) имеет RMSE равный 0.17, такое маленькое отклонение связано с тем, что входные данные нормированы между 0 и 1, так что, 0.17 мы можем рассматривать, как 17%. После этого мы проверили все степени вплоть до 15-ой и выяснили, что самое маленькое отклонение имеет полином 10-ой степени (**Рисунок 8**), а именно 11,5%, но взять мы решили полином 8-ой степени (**Рисунок 9**) с отклонением 12% и предсказывать по нему, так как, предсказывая будущее, больше подходит 8-ая степень, это прекрасно видно на графиках ниже.



**Рисунок 7.**

****

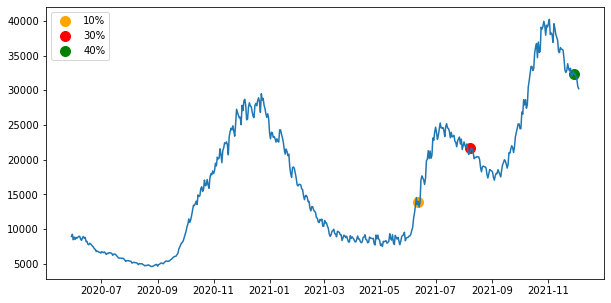
**Рисунок 8. Рисунок 9.**

Так как времени оставалось совсем мало, полностью нами задуманное реализовать не получилось, а именно, мы не использовали коррелирующие страны, чтобы уменьшить RMSE и сделать данные более точными.

**Статистистические данные о вакцинации:**

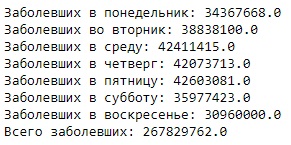
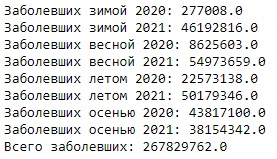
Так как у нас уже была вся информация, мы решили подвести некоторые интересные статистические данные.

Мы решили узнать, в какой момент в России вакцинировалось 10%, в какой 30% и в какой 40% человек (**Рисунок 8**). Узнать мы это смогли, благодаря тому же открытому ресурсу [Our World In Data](https://ourworldindata.org/), благодаря которому мы узнали и о заболеваемости.



**Рисунок 8.**

А также, мы можем узнать сезонность и частоту заболеваемости по дням недели и месяцам, что мы и сделали. И вот, что мы получили (приведённые данные актуальны на момент 7-го декабря 2021-го года):



**Итоги:**

Итак, подводя итоги нашей работы, можно сказать следующее: прежде всего, гипотеза наша подтвердилась, но предсказать заболеваемость мы смогли не так точно, как хотелось бы, кроме этого, был создан web-ресурс для визуализации данных и наглядной оценки, также, была найдена корреляцию стран между собой и подведена небольшая статистика.

**Перспективы:**

Мы и в будущем будем продолжать работу над нашим проектом, так как мы всё ещё можем уменьшить RMSE и сделать более точное предсказание. Кроме этого, мы можем точно утверждать, что наш проект помог нам развиться и дал много информации в сфере работы с данными и их обработке.