WEB-РЕСУРС ОБРАБОТКИ, ВИЗУАЛИЗАЦИИ И ПРЕДСКАЗАНИЯ СТАТИСТИКИ ЗАБОЛЕВАЕМОСТИ И СМЕРТНОСТИ ОТ COVID-19.

РАБОТУ ВЫПОЛНИЛИ:
БЕКТЕЕВ ГЕОРГИЙ АЛЕКСАНДРОВИЧ
БОРИСОВ ИГНАТ НИКОЛАЕВИЧ
НАУЧНЫЙ РУКОВОДИТЕЛЬ:
ШАМСУТДИНОВА НАТАЛЬЯ ВАЛЕРЬЕВНА

Проблематика

Правильное прогнозирование - это очень важная часть моделирования нынешней ситуации в любой стране. Своевременное прогнозирование появления новых штаммов вируса или новой волны может существенно облегчить работу первичного звена помощи и значительно уменьшить смертность. И, чтобы попробовать решить эту проблемы, мы и начали свою работу.

Гипотеза

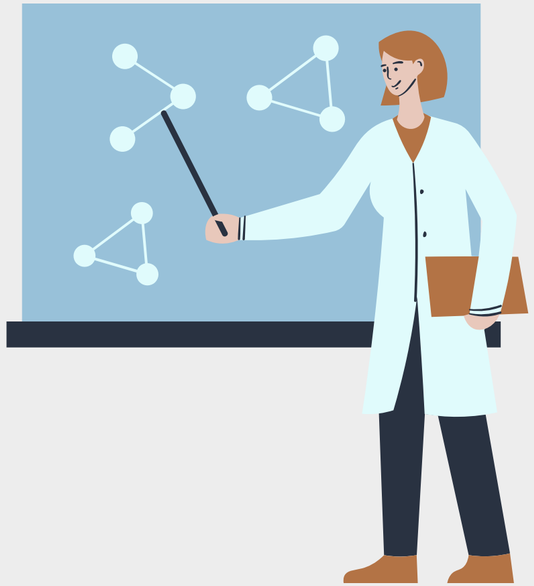
Имея на руках данные, которые дают нам информацию про заболеваемость и смертность граждан, почти во всех странах мира, можно вычислять коэффициент корреляции между ними и использовать его для поиска стран с самым схожим отнормированным ростом и спадом заболеваемости, что далее можно использовать для обучения модели регрессии и предсказывать заболеваемость на 2-3 дня вперёд

Сравнение с аналогами

Яндекс

ЦЕЛЬ

Создать доступный для всех сервис для просмотра статистики заболеваемости и смертности от covid-19 с предсказанием на несколько дней вперёд.



ЗАДАЧИ

01

Визуализация смертности и заболеваемости в разных странах

02

Поиск и сравнение стран с близкой ситуацией заболеваемости

03

Прогноз количества заболевших на несколько дней вперёд

ПЛАНИРОВАНИЕ

- Сбор и анализ данных
- Преобразования данных
- Визуализация данных
- Нормирование
- Коэффициент корреляции
- Линейная регрессия и первые прогнозы
- Анализ временных рядов
- Создание web-ресурса

Таблица со всеми данными

	iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	...
0	AFG	Asia	Afghanistan	2020-02-24	5.0	5.0	NaN	NaN	NaN	NaN	...
1	AFG	Asia	Afghanistan	2020-02-25	5.0	0.0	NaN	NaN	NaN	NaN	...
2	AFG	Asia	Afghanistan	2020-02-26	5.0	0.0	NaN	NaN	NaN	NaN	...
3	AFG	Asia	Afghanistan	2020-02-27	5.0	0.0	NaN	NaN	NaN	NaN	...
4	AFG	Asia	Afghanistan	2020-02-28	5.0	0.0	NaN	NaN	NaN	NaN	...
...	...	...	...	...	...	...	...	...	...	...	...
160694	ZWE	Africa	Zimbabwe	2022-02-05	230402.0	232.0	141.000	5362.0	5.0	4.143	...
160695	ZWE	Africa	Zimbabwe	2022-02-06	230402.0	0.0	134.571	5362.0	0.0	3.571	...
160696	ZWE	Africa	Zimbabwe	2022-02-07	230402.0	0.0	105.143	5362.0	0.0	3.429	...
160697	ZWE	Africa	Zimbabwe	2022-02-08	230603.0	201.0	107.429	5366.0	4.0	2.286	...
160698	ZWE	Africa	Zimbabwe	2022-02-09	230740.0	137.0	104.000	5367.0	1.0	2.143	...

Сбор и анализ данных

Преобра-зование

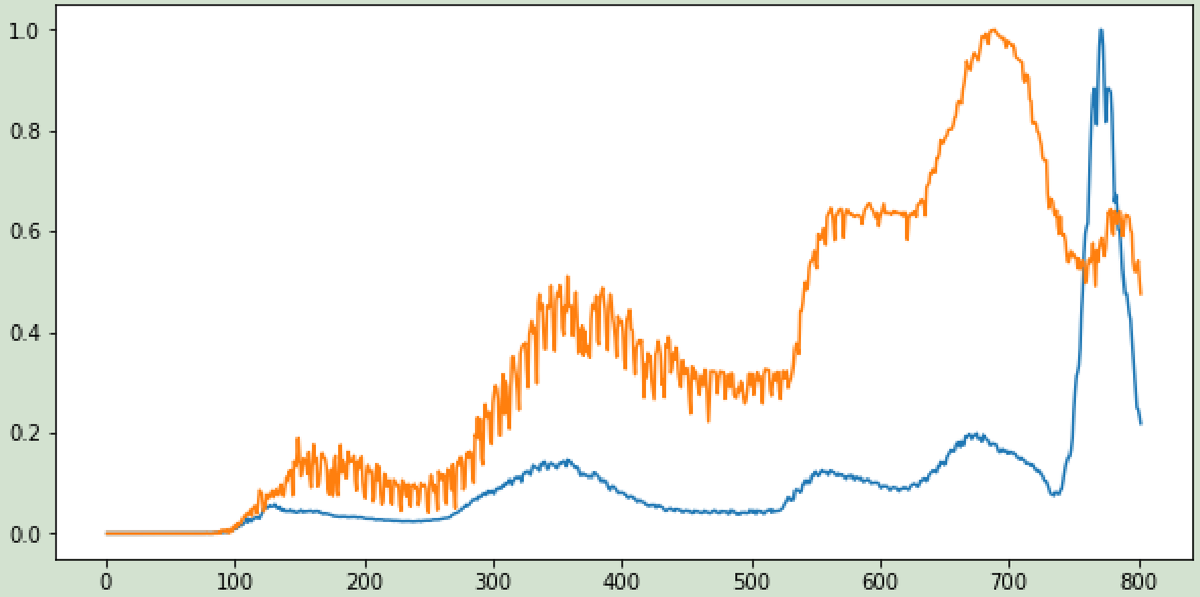
Преобразованная таблица смертности

	Afghanistan	Africa	Albania	Algeria	Angola	Argentina	Armenia	Asia	Australia	Austria	...	United Kingdom	United States	Upper middle income	Uruguay	Uzbekistan	Venezue
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	(
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	(
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	(
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	(
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	(
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
703	1.0	98.0	1.0	7.0	0.0	4.0	25.0	1080.0	8.0	60.0	...	127.0	507.0	2199.0	0.0	3.0	9
704	0.0	82.0	1.0	7.0	0.0	3.0	14.0	3996.0	6.0	43.0	...	54.0	161.0	2281.0	2.0	2.0	7
705	4.0	143.0	5.0	8.0	0.0	34.0	8.0	1238.0	9.0	48.0	...	46.0	1382.0	2632.0	2.0	3.0	(
706	1.0	128.0	7.0	3.0	0.0	23.0	27.0	1092.0	7.0	77.0	...	180.0	1609.0	2441.0	1.0	3.0	9
707	0.0	160.0	4.0	8.0	0.0	5.0	10.0	1347.0	10.0	58.0	...	163.0	1714.0	2559.0	1.0	1.0	6

Преобразованная таблица заболеваемости

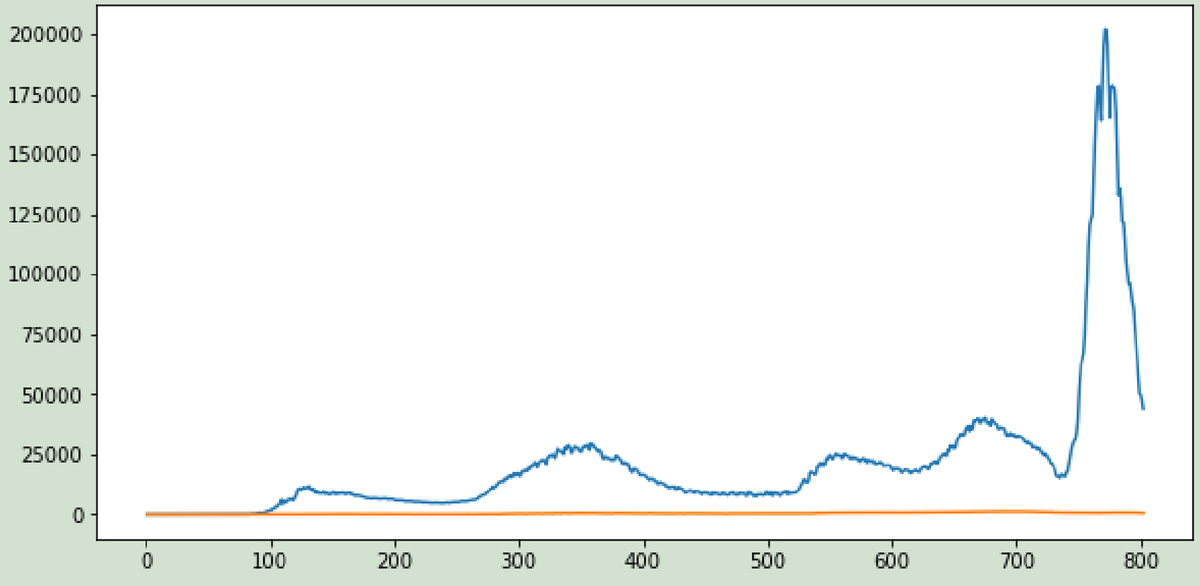
	Afghanistan	Africa	Albania	Algeria	Angola	Argentina	Armenia	Asia	Australia	Austria	...	United Kingdom	United States	Upper middle income	Uruguay	Uzbekistan	V
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
703	19.0	20535.0	357.0	185.0	21.0	1690.0	419.0	74300.0	1249.0	7304.0	...	41574.0	61013.0	120138.0	279.0	212.0	
704	14.0	14651.0	328.0	172.0	15.0	1294.0	327.0	89989.0	1285.0	5192.0	...	43361.0	34215.0	109532.0	104.0	228.0	
705	54.0	11032.0	172.0	193.0	0.0	2477.0	95.0	74170.0	1434.0	4625.0	...	51746.0	192917.0	98156.0	178.0	223.0	
706	9.0	20055.0	393.0	197.0	42.0	3089.0	240.0	83308.0	1705.0	4233.0	...	45473.0	108930.0	117084.0	237.0	147.0	
707	34.0	33673.0	346.0	188.0	31.0	1881.0	410.0	89740.0	1654.0	5663.0	...	51003.0	151739.0	127037.0	300.0	201.0	

Визуализация

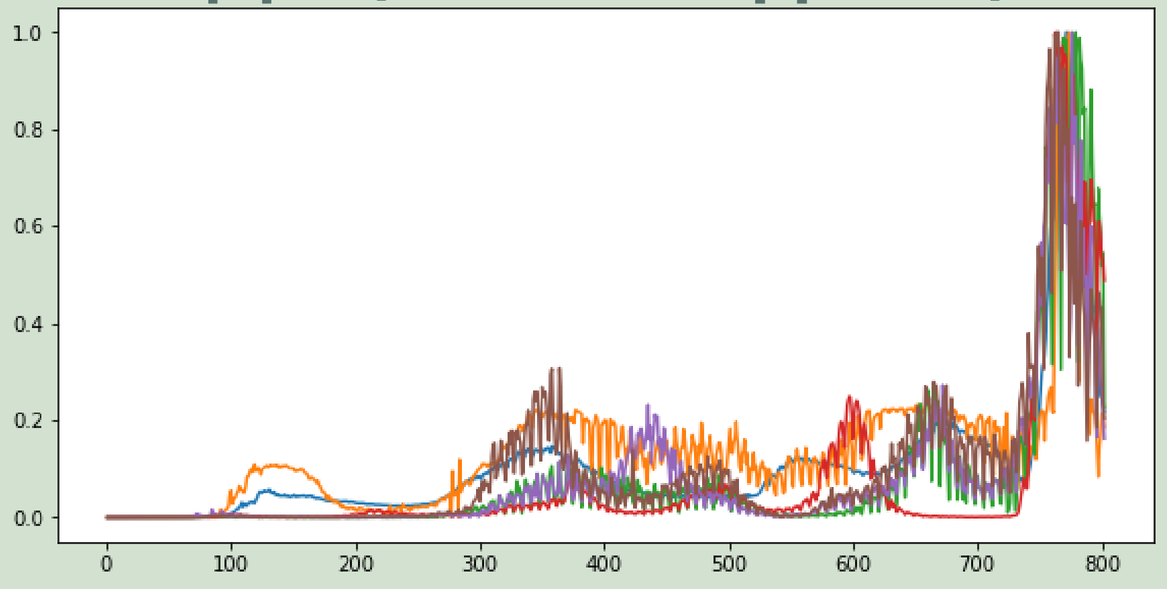


Пронормированные данные

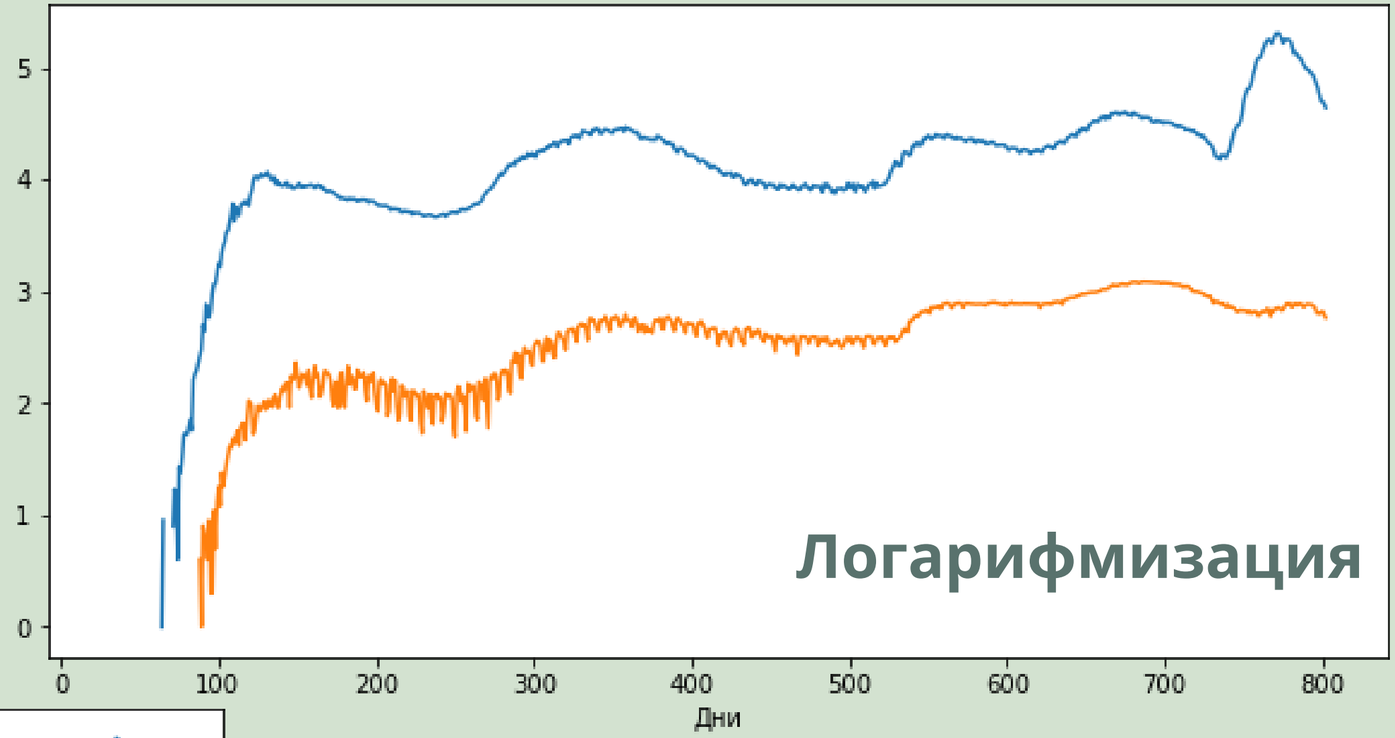
Исходные данные



5 стран с близким коэффициентом корреляции



Нормирование



$$\frac{X - Xmin}{Xmax - Xmin}$$

Формула линейного нормирования

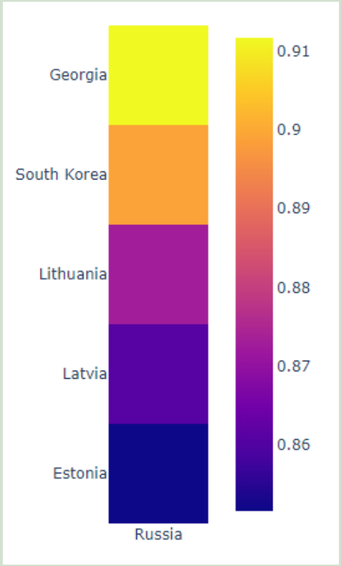
$$log_{10}(x)$$

Формула логорифмизации

Коэффициент корреляции

	Afghanistan	Africa	Albania	Algeria	Angola	Argentina	Armenia	Asia	Australia	Austria	...	United Kingdom	United States	Upper middle income	Uru
Afghanistan	1.000000	0.411252	-0.302237	0.267160	0.097353	0.456442	-0.173736	0.161754	-0.195796	-0.211170	...	-0.004316	-0.187491	0.289402	0.33
Africa	0.411252	1.000000	0.205488	0.698574	0.315073	0.397455	0.009542	0.408765	0.027876	-0.052037	...	0.521509	0.472213	0.705777	0.11
Albania	-0.302237	0.205488	1.000000	0.083601	0.226787	-0.108285	0.384127	0.070557	0.350732	0.432185	...	0.499081	0.648980	0.427375	-0.06
Algeria	0.267160	0.698574	0.083601	1.000000	0.162005	0.265805	0.095269	0.310786	-0.083901	0.026840	...	0.275045	0.367092	0.477925	-0.07
Angola	0.097353	0.315073	0.226787	0.162005	1.000000	0.399841	0.308476	0.532233	0.375581	0.054799	...	0.305158	0.268892	0.490564	0.27
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
Venezuela	0.222709	0.339064	0.122657	0.189389	0.478504	0.478547	0.163450	0.578020	0.337436	0.136347	...	0.252199	0.086905	0.587856	0.41
World	0.054797	0.589719	0.504144	0.418365	0.495020	0.561981	0.415244	0.745674	0.211208	0.451381	...	0.569977	0.687301	0.898337	0.45
Zambia	0.671206	0.595082	0.056958	0.252849	0.075059	0.434829	-0.235258	0.110901	-0.181217	-0.158544	...	0.210271	0.027054	0.436341	0.27
Zimbabwe	0.309804	0.627357	-0.037683	0.455650	0.089417	0.190109	-0.122323	0.164572	0.056688	-0.029277	...	0.421974	0.148718	0.331836	-0.02
nu	0.198259	0.513468	0.465049	0.284306	0.501768	0.367914	0.358423	0.610565	0.597240	0.491895	...	0.692797	0.394718	0.797236	0.34

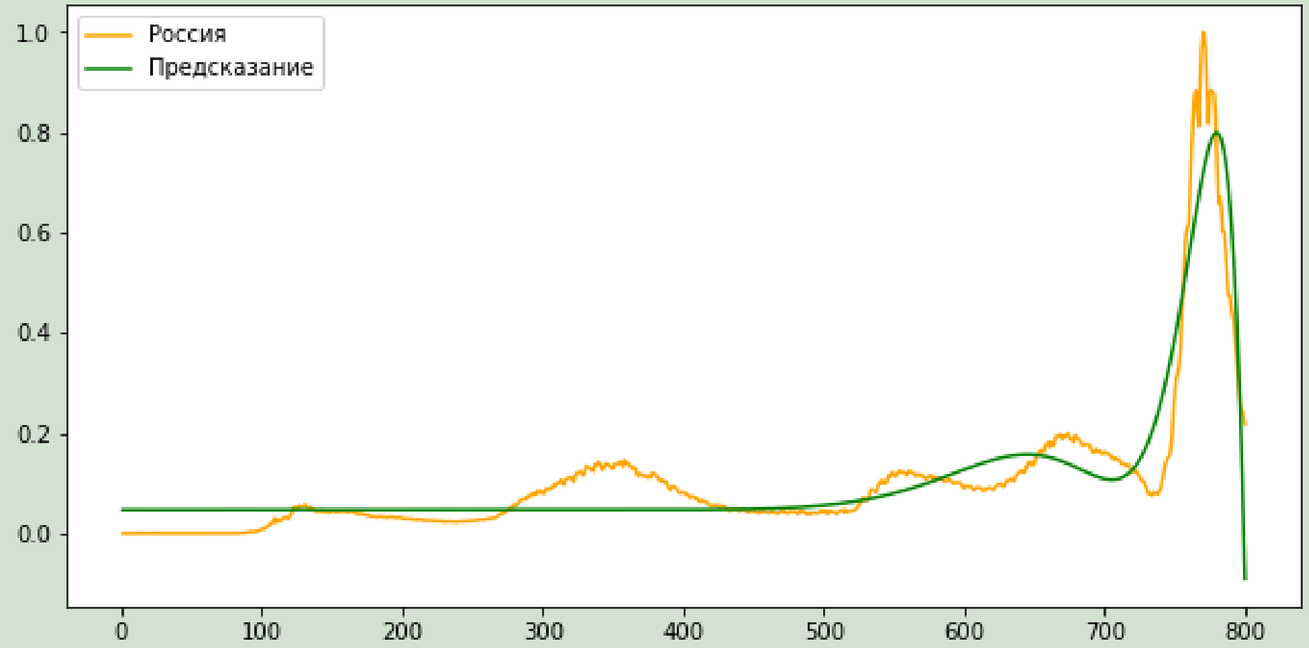
Таблица с коррелированными данными о смертности



Тепловая карта

$$r_{xy} = \frac{\Sigma(x_i - \underline{x}) \times (y_i - \underline{y})}{\sqrt{\Sigma(x_i - \underline{x})^2 \times \Sigma(y_i - \underline{y})^2}}$$

Формула коэффициента корреляции



$$y = f(x, b) + \varepsilon,\ E(\varepsilon)$$

Формула линейной регрессии

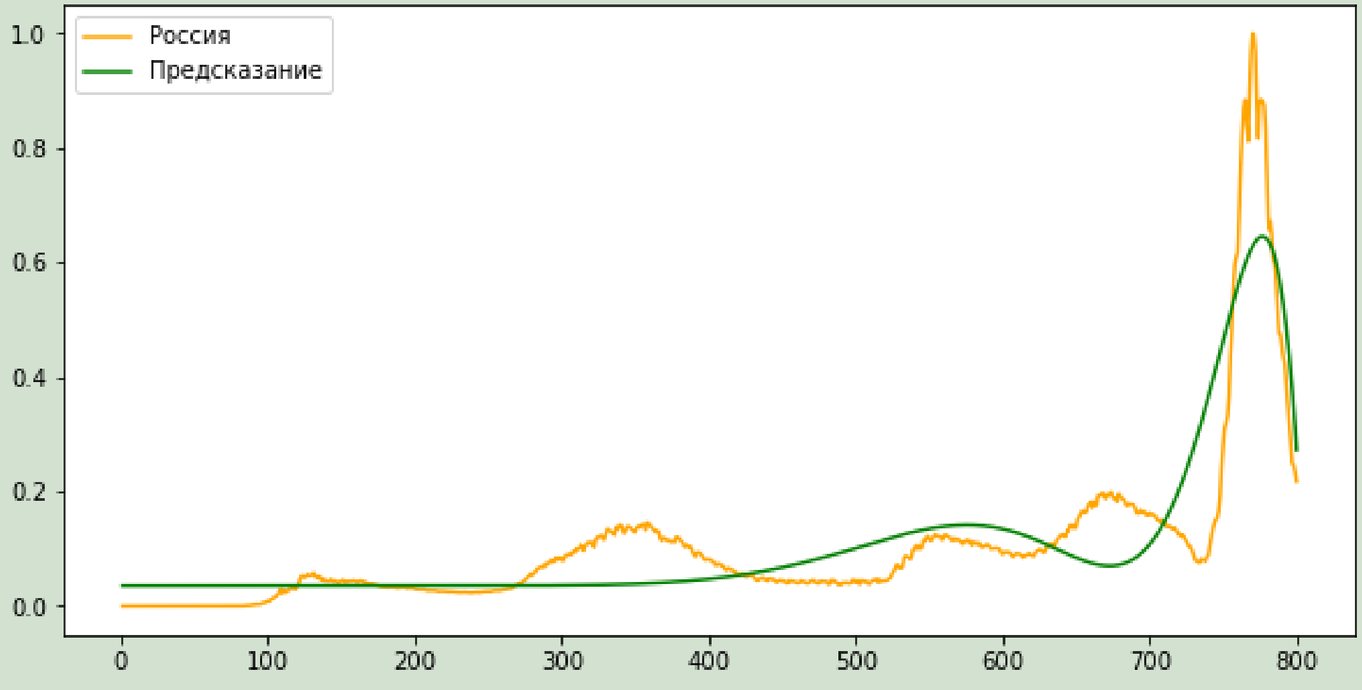
RMSE полинома 24-ой степени - 5%

RMSE полинома 16-ой степени - 7%

Полином 24-ой степени

Полином 16-ой степени

Обучение регрессии

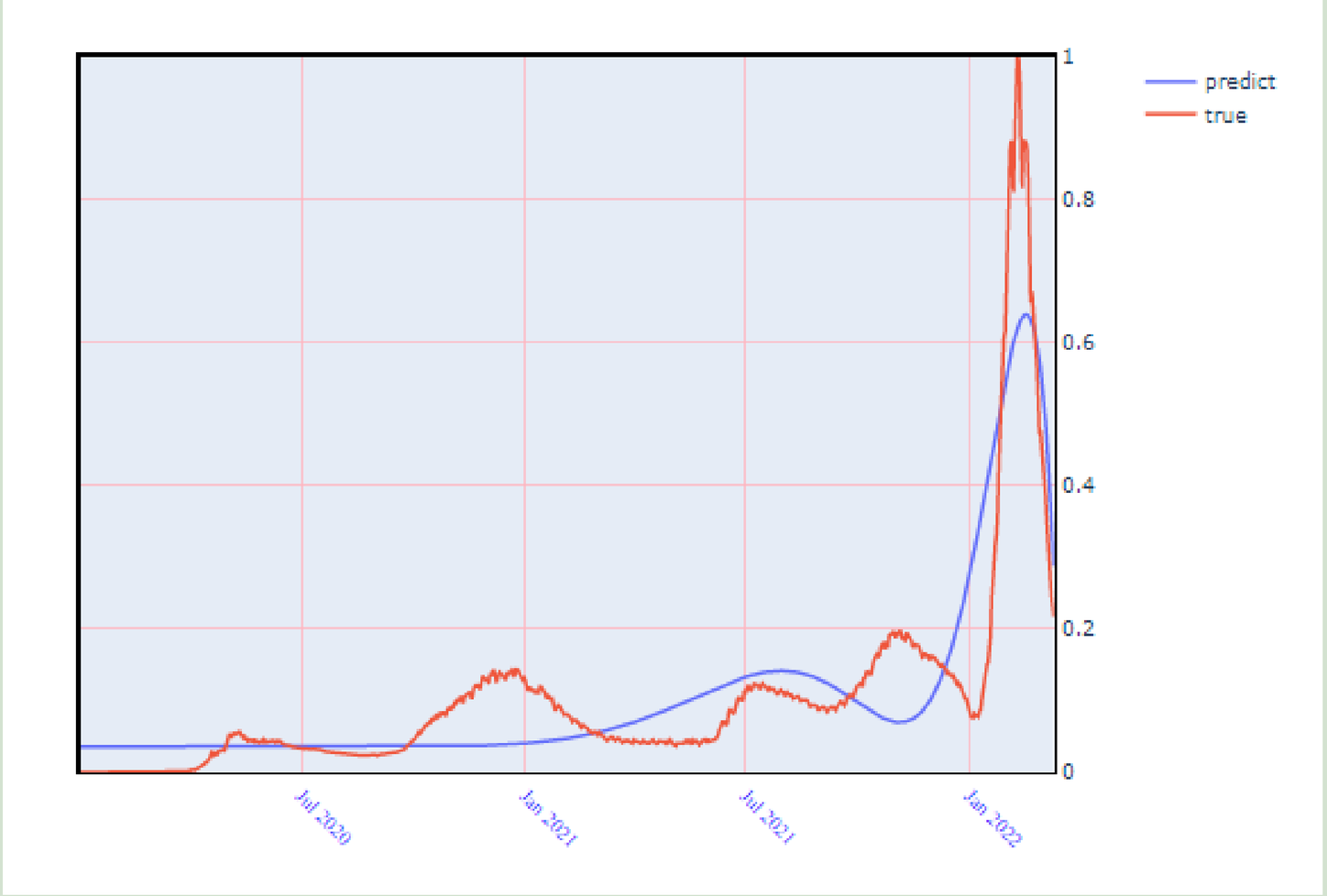


14 марта: 117727 ± 10595

15 марта: 117727 ± 10595

16 марта: 117677 ± 10591

Первые прогнозы



Предсказание с сайта

$$N \times (x_{max} - x_{min}) + x_{min},$$

где N - это предсказанное число

Формула, по которой мы выражали предсказанное из пронормированного числа

Анализ временных рядов

Метод полиномиальной регрессии предсказал:

На 14 марта: 34992
На 15 марта: 34312
На 16 марта: 33603

RMSE - 1%, полином 13-ой степени

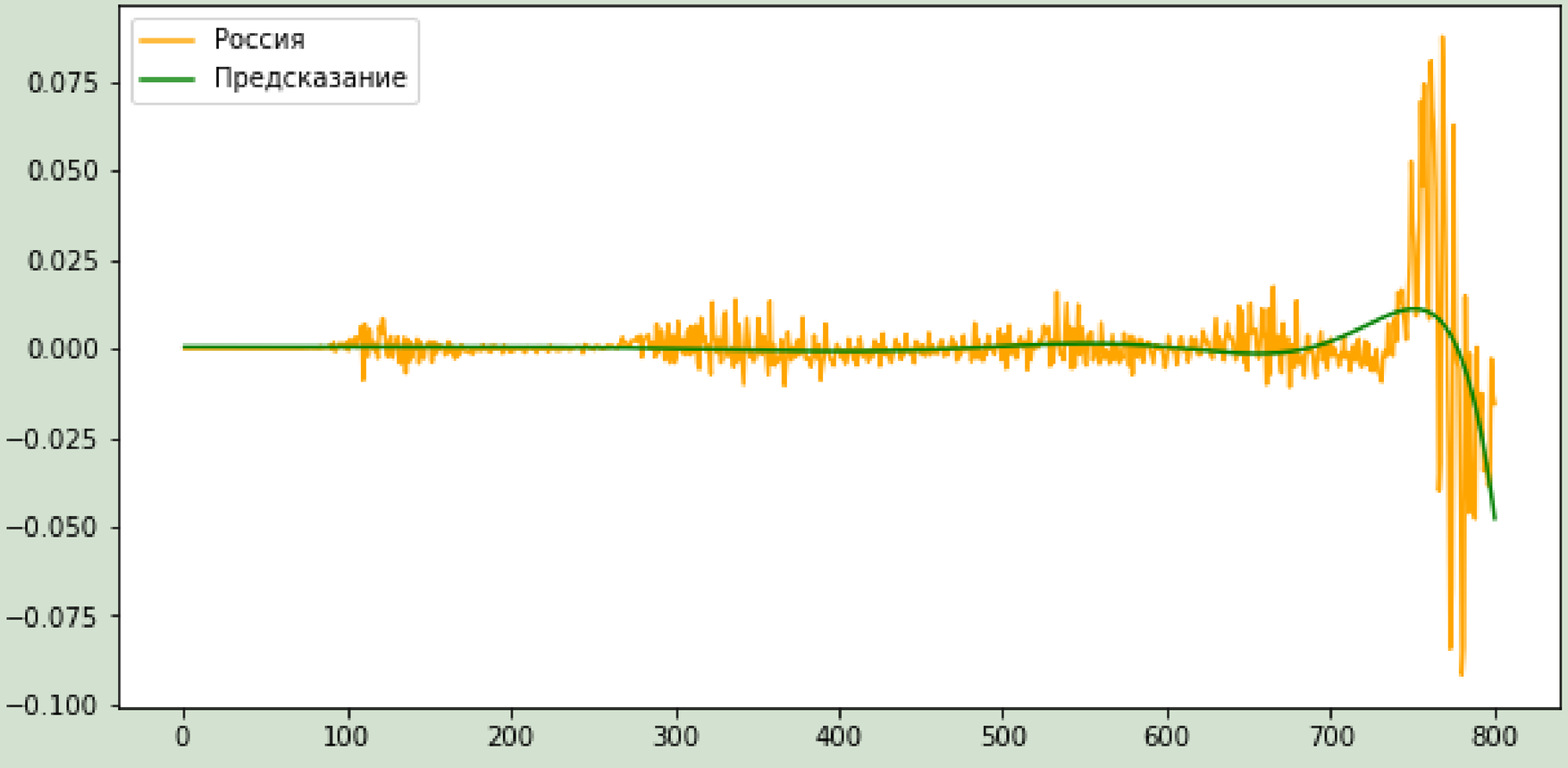
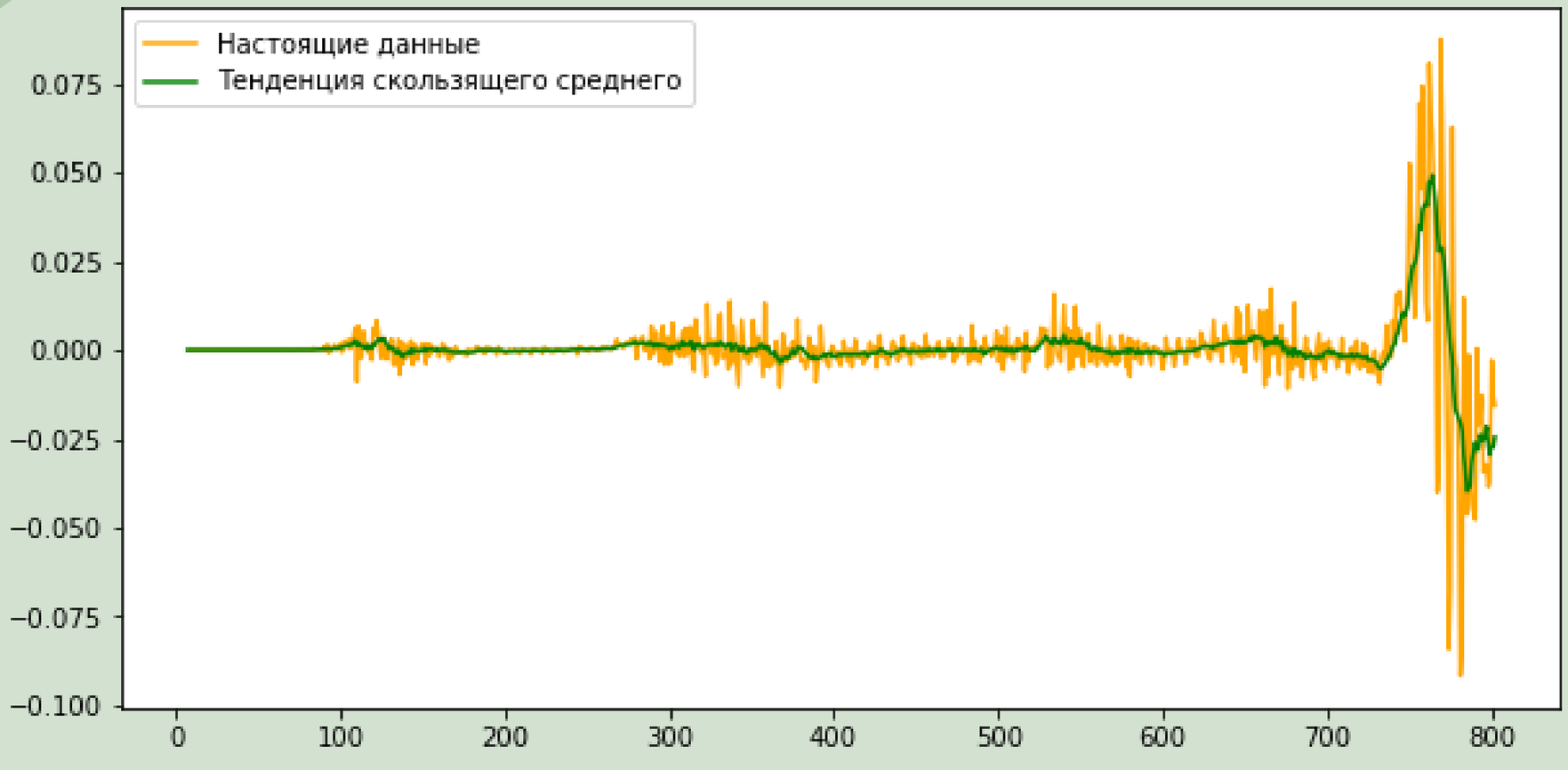


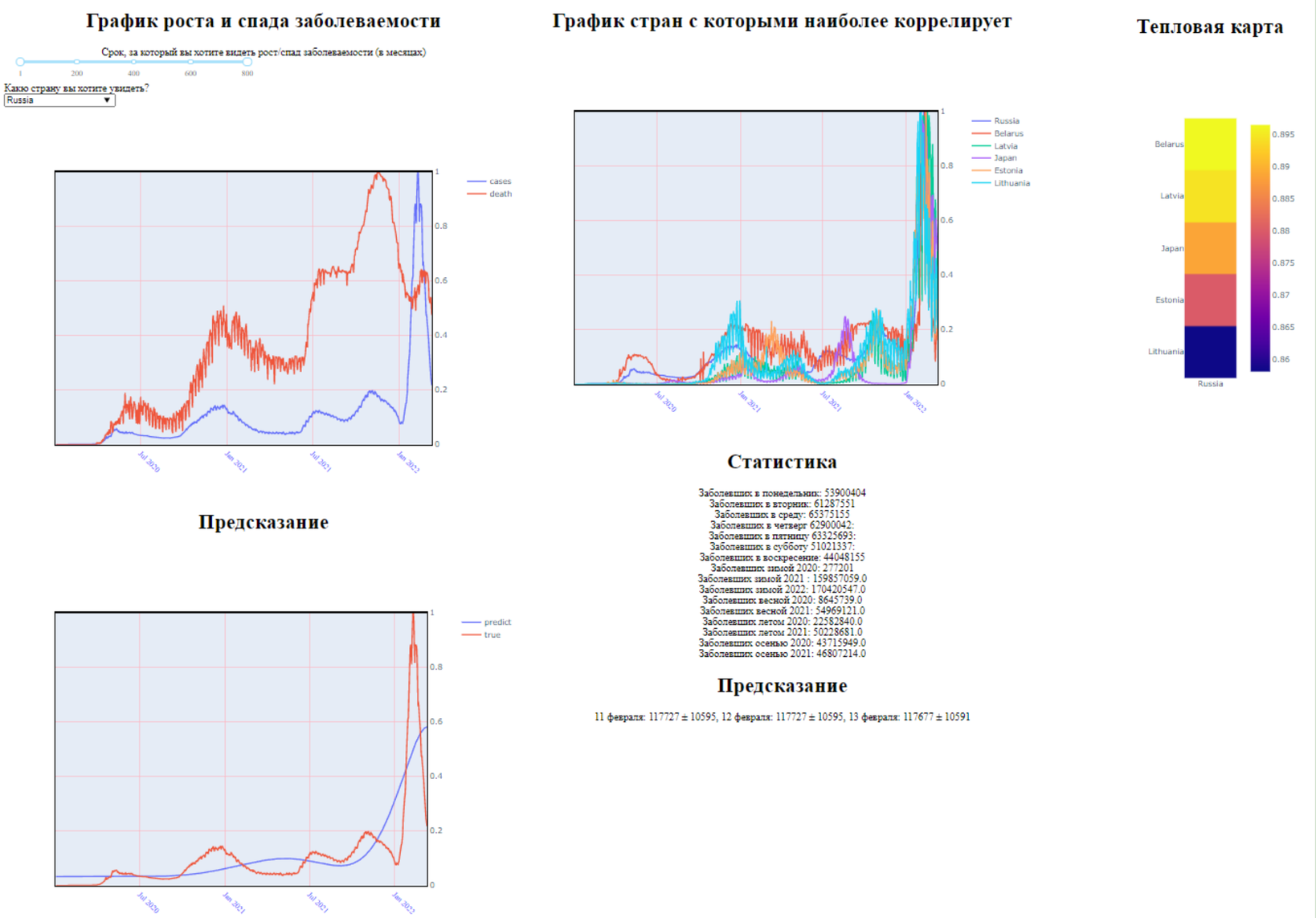
График строится по разнице заболевших между соседними днями

Анализ временных рядов

Тенденция скользящего среднего: Предсказание по 7 последним дням, так как средний латентный период - 4 дня, а инкубационный - 3 дня

На 14 марта: 40726
На 15 марта: 37635
На 16 марта: 34621





Скриншот нашего сайта

Создание web-ресурса

Итоги нашей работы:

- Алгоритм обработки данных заболеваемости covid-19
- Визуализация данных
- Статистическая обработка данных (сезонные изменения)
- Модели прогнозирования и их оценка
- Web-сайт с доступом к информации


THANKS