



# Прогнозирование развития COVID-19 в Италии в 2021 г. Covid-19 forecasting for Italy in 2021 real time data.



Итоговая аттестация.

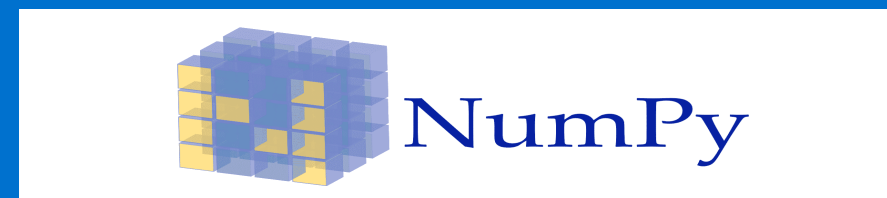
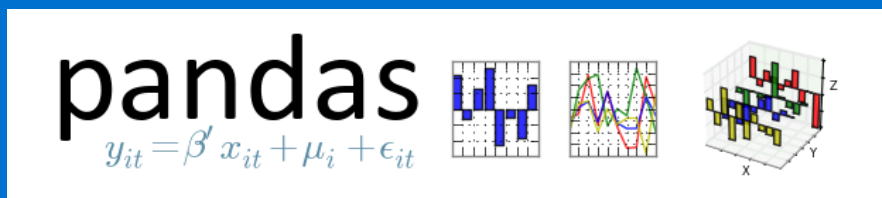
Выполнила: Никулина М.К.

Ссылка на репозиторий: [GitHub](#)

Курс:

«Аналитик: искусство управлять данными»

INNOPOLIS  
UNIVERSITY



$$SARIMA \underbrace{(p, d, q)}_{non-seasonal} \underbrace{(P, D, Q)_m}_{seasonal}$$

**Цель:** провести анализ данных о заболеваемости COVID-19 на примере Италии, предложить и настроить прогностическую модель, выполнить прогноз и сравнить результаты с новой статистикой (полученной после 5 апреля 2021 года).

### Задачи :

- Определение проблемы;
- Сбор информации;
- Предварительный (разведочный) анализ.
- Выбор и настройка моделей;
- Использование и оценка модели прогнозирования.



Подготовка к анализу:

- Импорт библиотек и функций.
- Импортируем модели и метрики

Загружаем данные:

- Ознакомление с данными
- Выбираем страну ( Италия )

Предобработка данных:

- Удаление лишних столбцов.
- Фильтрация и удаление пропусков. ( для сохранения целости временного ряда.)

Разведочный анализ:

- Выводим общую статистику по итоговым столбцам.
- Строим графическое отображение.
- Выводим ETS декомпозицию

Выбор и настройка моделей:

- SARIMA
- Хольта-Винтерса
- PROPHET

Вывод.

## План работы:






- COVID-19 - это вызывающий заболевание штамм коронавируса, появившийся в декабре 2019 года и приведший к продолжающейся глобальной пандемии. Возможность предвидеть путь пандемии имеет решающее значение. Это важно для того, чтобы определить, как бороться, и отследить его распространение.
- Возьмем для исследования общедоступные ежедневные данные о COVID-19 по странам за период с 24 марта 2020 года по 19 декабря 2021г.

<https://github.com/owid/covid-19-data/tree/master/public/data>

## Данные о COVID-19 (коронавирус) от журнала *Our World in Data*

 Загрузите наш полный набор данных о COVID-19: [CSV](#) | [XLSX](#) | [JSON](#)

Наш полный набор данных о COVID-19 - это набор данных о COVID-19, которые хранятся в журнале «*Наш мир в данных*». Мы будем обновлять его ежедневно на протяжении пандемии COVID-19. Он включает следующие данные:

Метрики	Источник	Обновлено	Страны
Прививки	Официальные данные собраны командой "Наш мир в данных"	Ежедневно	218
Тесты и позитив	Официальные данные собраны командой "Наш мир в данных"	Еженедельно	140
Больница и ОИТ	Официальные данные собраны командой "Наш мир в данных"	Еженедельно	39
Подтвержденные случаи [заболевания]	Данные JHU CSSE COVID-19	Ежедневно	217
Подтвержденные смерти	Данные JHU CSSE COVID-19	Ежедневно	217
Скорость воспроизводства	Арройо-Мариоли Ф, Буллано Ф, Кучинскас С, Рондон-Морено С	Ежедневно	185
Ответы политики	Оксфордский трекер реакции правительства на COVID-19	Ежедневно	186
Другие интересные переменные	Международные организации (ООН, Всемирный банк, ОЭСР, ИМЕ...)	Фиксированный	241

# Предобработка данных:

Нам даны 67 колонок данных по 160 странам. Для проведения анализа данных и прогноза ситуации по covid-19 - сделаем выборку данных по одной стране (Италия)

**Уберем малоинформативные колонки данных, и оставим 5 наиболее важных для анализа и построения моделей прогнозирования:**

- дата
- общее количество случаев
- новые случаи заболеваний
- общее количество смертности
- новые случаи смертности

**Уберем пропуски данных.**

**Меняем тип данных ( для коррективного считывания данных.)**

	date	total_cases	new_cases	total_deaths	new_deaths
67484	2020-01-31	2.0	2.0	0.0	0.0
67485	2020-02-01	2.0	0.0	0.0	0.0
67486	2020-02-02	2.0	0.0	0.0	0.0
67487	2020-02-03	2.0	0.0	0.0	0.0
67488	2020-02-04	2.0	0.0	0.0	0.0
...	...	...	...	...	...
68166	2021-12-13	5238221.0	12704.0	134929.0	98.0
68167	2021-12-14	5258886.0	20665.0	135049.0	120.0
68168	2021-12-15	5282076.0	23190.0	135178.0	129.0
68169	2021-12-16	5308180.0	26104.0	135301.0	123.0
68170	2021-12-17	5336795.0	28615.0	135421.0	120.0

687 rows x 5 columns



# Общий график Covid 19 - Италия

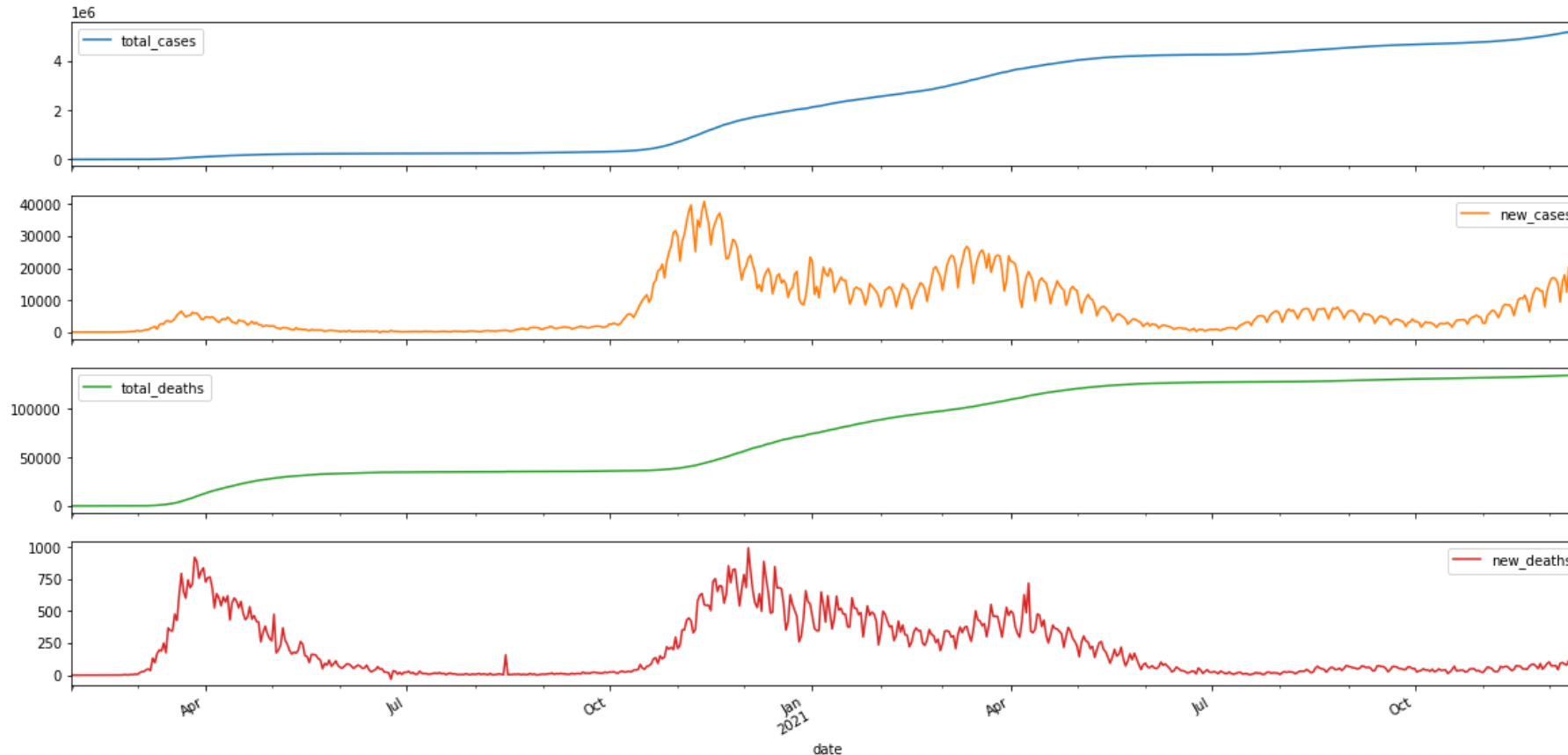
## Метрики для прогнозирования:

- 1) total\_cases - накопительный итогом кол-во новых случаев заболеваний
- 2) new\_cases - новые случаи заболеваний за день
- 3) total\_deaths - накопительный итогом кол-во смертей
- 4) new\_deaths - новые случаи смертей за день

```
df1[['total_cases', 'new_cases', 'total_deaths', 'new_deaths']].plot(subplots=True, figsize=(20, 10), title = 'COVID-19 Италия');
```



COVID-19 Италия



Начало активного роста заболеваний начинает прослеживаться с конца октября 2020 года, рост продолжился до конца 2021 года, приближаясь к выходу на плато

Что явно просматривается на графике новых случаев заболевания в день.

Общий график количества смертей показывает нам волнообразный рост с середины марта 2020

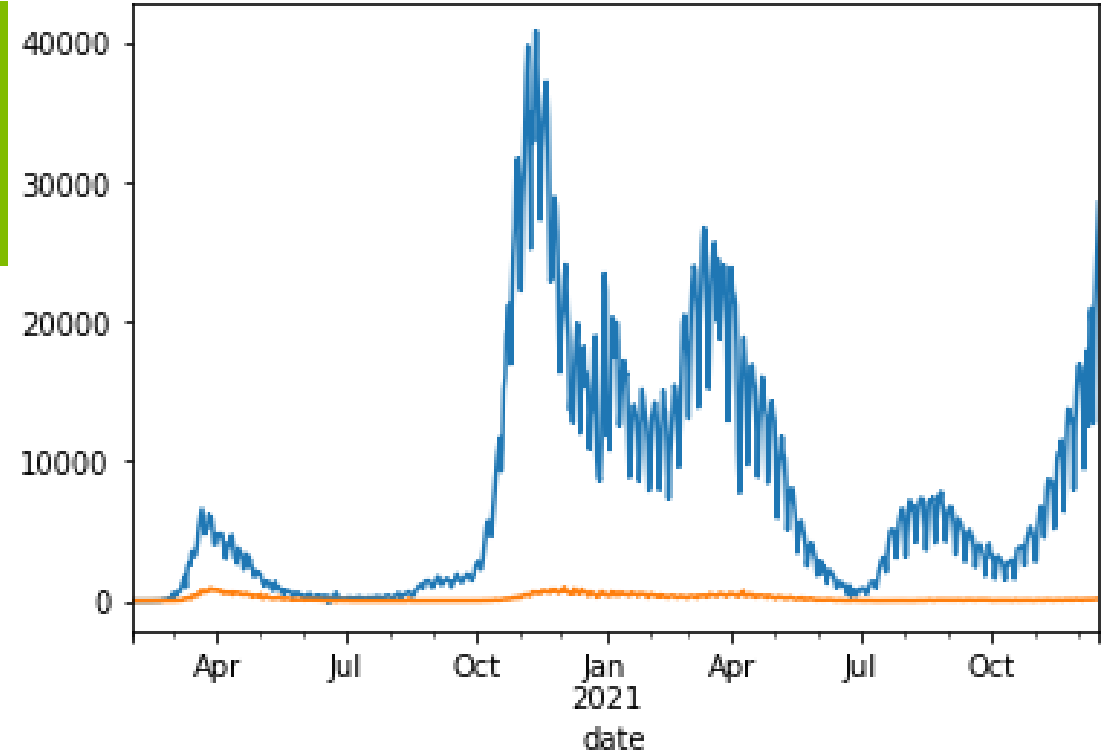
По ежедневной динамике новых смертей видим три пика, в апреле и декабре 2020, последний всплеск в апреле 2021, с июля 2021 года уровень стабильный

# Объект анализа - Италия

Построим графики новых случаев заболеваний и новых смертельных случаев, посмотрим на их корреляцию. На пересечении new\_cases и new\_deaths 0.61 наблюдается корреляция близкая к линейной, зависимость двух признаков. Говоря нам о том, что наблюдается тесная связь признаков.

```
df1.corr()
```

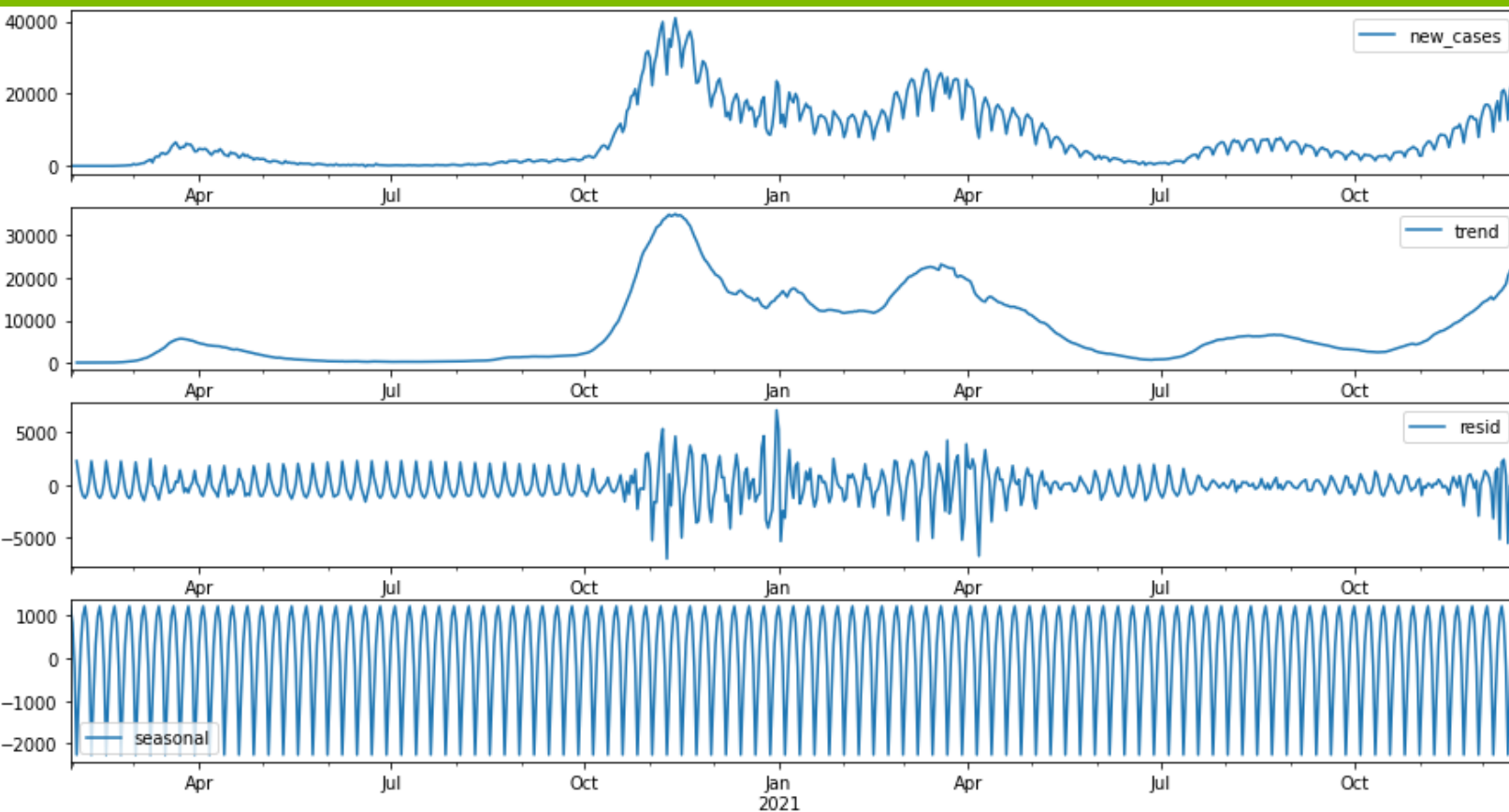
	total_cases	new_cases	total_deaths	new_deaths
total_cases	1.000000	0.190604	0.986076	-0.133710
new_cases	0.190604	1.000000	0.182424	0.612691
total_deaths	0.986076	0.182424	1.000000	-0.151847
new_deaths	-0.133710	0.612691	-0.151847	1.000000





# Выполним ETS декомпозицию, используя аддитивную модель ('additive').

Seasonal\_decompose в увеличенном виде  
Видим, что в данных имеется сезонность



Поскольку имеется четко выраженная сезонность, это дает нам возможность использовать модели прогнозирования временных рядов:

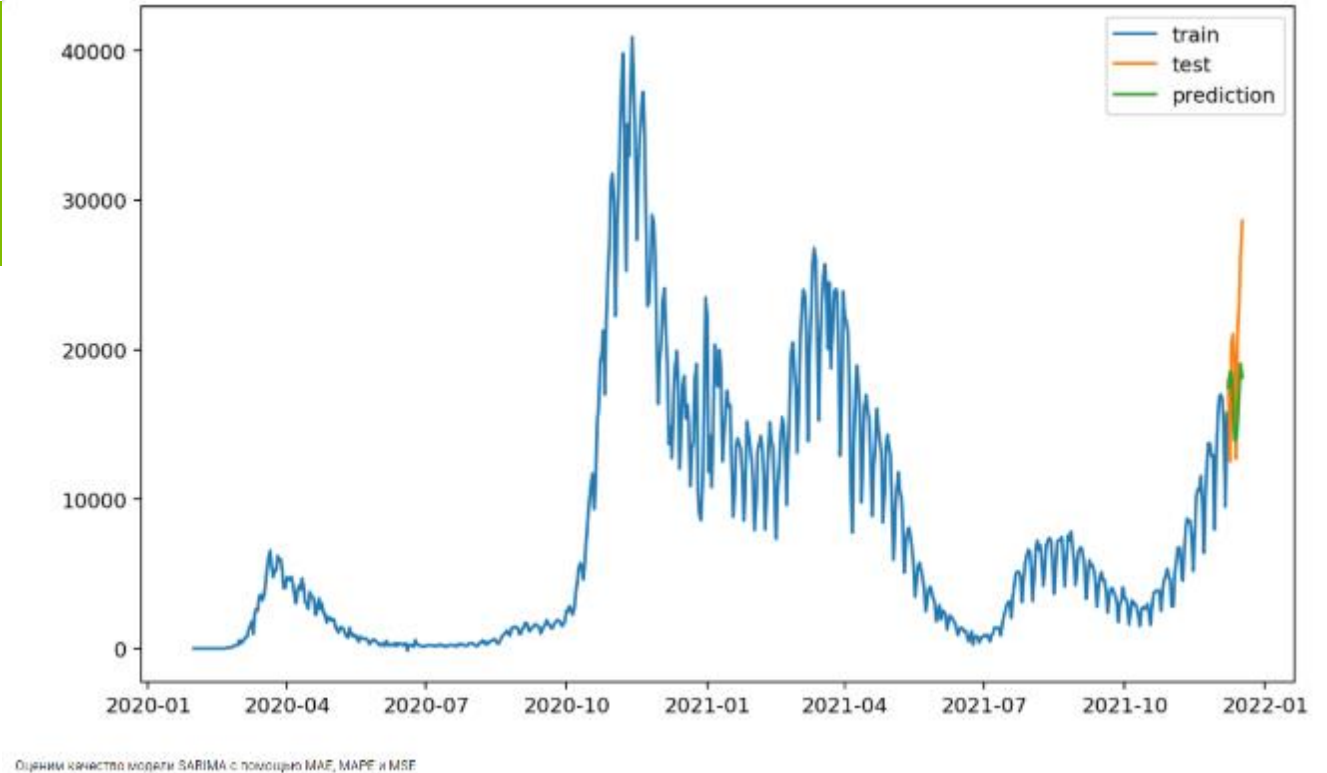
- SARIMA
- Хольта-Винтерса
- PROPHET

Будем выстраивать прогноз на 10 дней.

# 1й метод прогнозирования - SARIMAX

Модель сезонного авторегрессионного скользящего среднего с учетом сезонности. Используем функцию `auto_arima`

Результат прогнозирования на 10 дней подтверждает, что волна новых заболеваний и случаев смертности ещё не закончилась. Говоря нам о том, что прогноз достаточно точный.

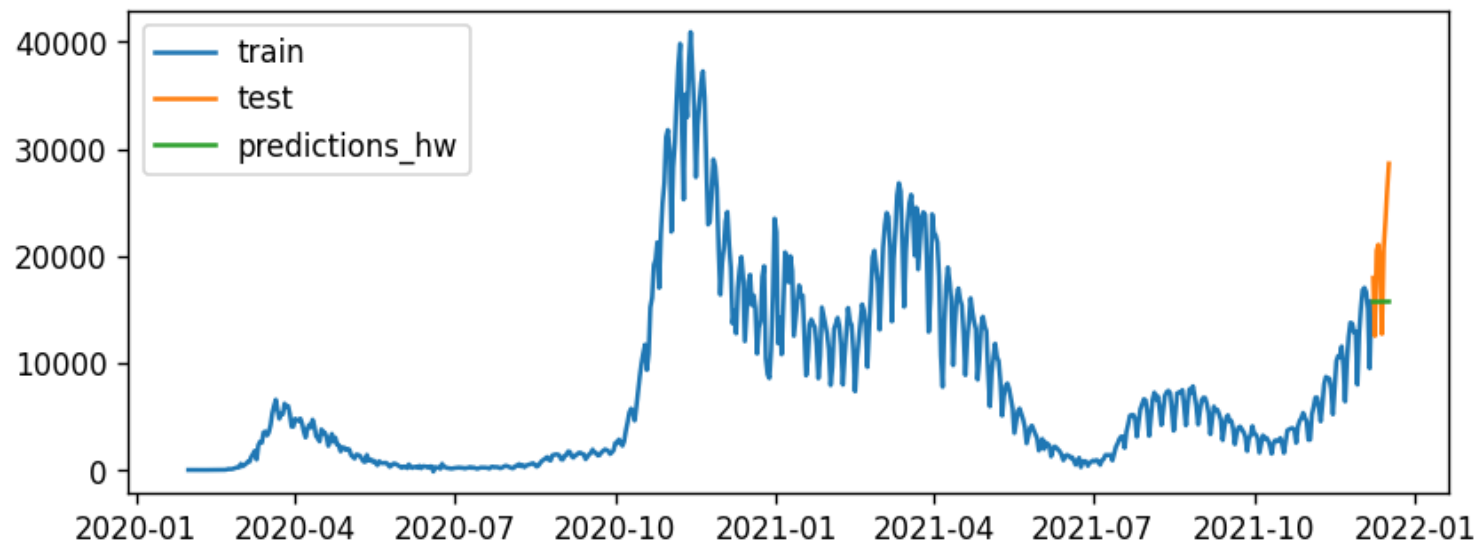


MSE Error: 32124489.87  
RMSE Error: 5667.847023

# 2й метод прогнозирования - Экспоненциальное сглаживание - модель Хольта-Винтера (Holt-Winters' Model)

Метод Хольта-Винтерса – это трехпараметрическая модель прогноза, которая учитывает: – сглаженный экспоненциальный ряд; – тренд; – сезонность.

Модель показала большую погрешность прогнозирования, не стоит опираться на эти данные.



Рассчет точности полученного прогноза

MAE: 5772.772040200725

MAPE: 0.2695049734802351

MSE: 44180412.21856908

RMSE: 6646.834752

**Вывод:** Модель Хольта-Винтера показала большую погрешность прогнозирования среднеквадратичная ошибка составила 6646, от исходных данных.

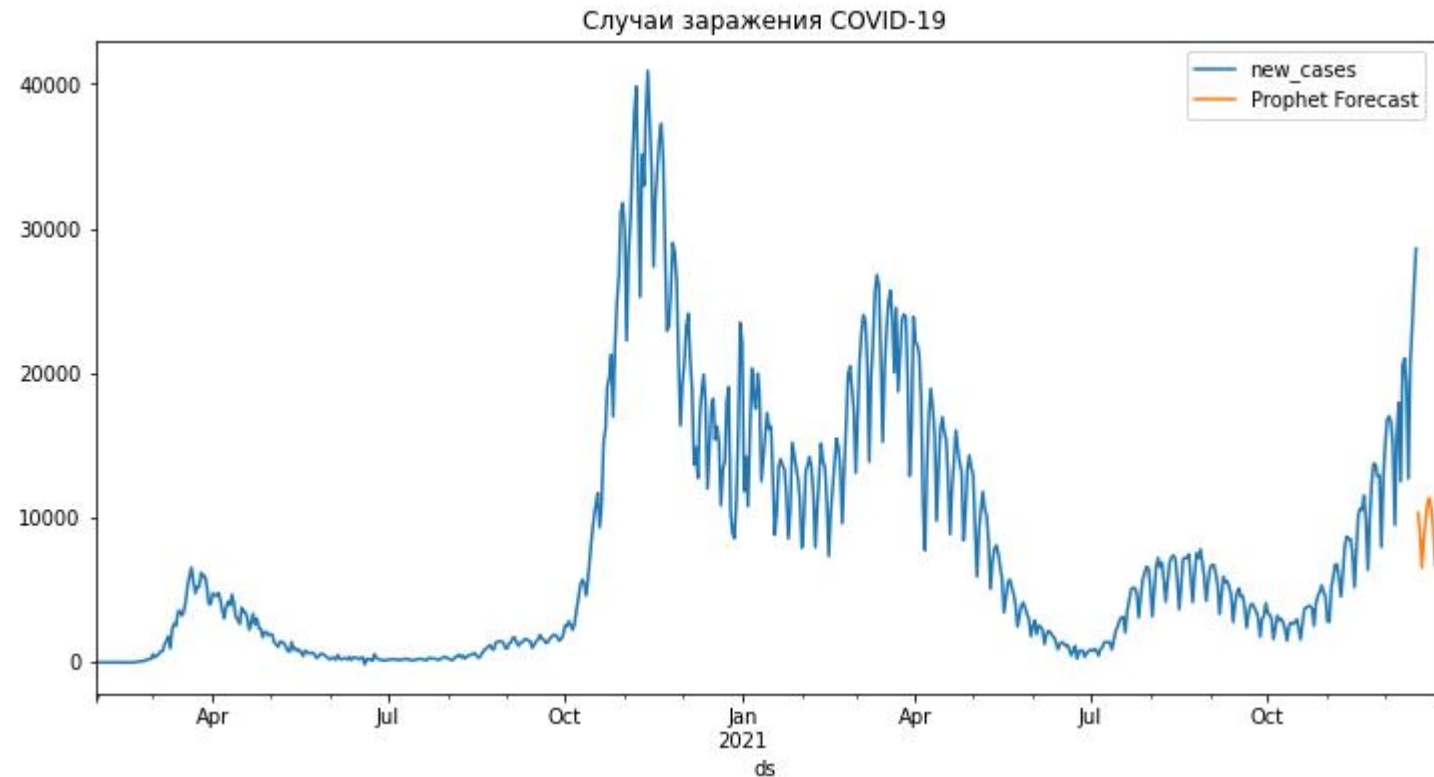
## 3-й метод прогнозирования - Prophet

При построении модели не была найдена годовая и ежедневная сезонность, зато была установлена еженедельная сезонность, что подтверждает наши выводы, полученные ранее.

### Выводы работы метода PROPHET:

Мы создали модель с подобранными параметрами, обучили модель на обучающей выборке данных. Сделали прогноз на 10 дней вперед. Модель прогнозирует о снижении кол-ва заболеваний, что не точно отображает действительность.

Prophet MAE Error: 13011.13948  
Prophet MSE Error: 188573579.8  
Prophet RMSE Error: 13732.20958  
Prophet MAPE Error: 62.57596306



# Вывод:

Для построения прогнозной модели применены следующие методы:

- SARIMAX
- Хольта-Винтерса
- Prophet;

Была проведена предобработка и подготовка данных.

Оценка качества моделей прогнозирования новых случаев заболеваемости COVID - 19 произвели с помощью ключевых показателей - метрик. Оценки качества была определена

- средняя абсолютная ошибка в процентах (MAPE)
- средняя абсолютная разница между предсказаниями и фактическими значениями (MAE);
- среднеквадратичная ошибка (MSE);
- среднеквадратичной ошибкой (RMSE). Данная система метрик позволит объективно оценить точность моделей на тестовой выборке и сравнить используемые методы.**

Самые точные показатели у модели Sarima.

Для прогноза лучше всего использовать эту модель.

Но в целом все модели показали достаточно большую погрешность.



INNOPOLIS  
UNIVERSITY



Спасибо за внимание!