**Отчёт о разработке интерактивного виджета для предсказания эпидемиологической ситуации в странах мира Агапова Никиты Александровича**

Предоставлен экспертам соревнований World Skills Russia за период с «7» апреля 2021 года по «8» апреля 2021 года

На отчётный период передо мной стояли следующие задачи

1. 7 апреля загрузить данные с открытого источника и проанализировать их
2. Выполнить предобработку данных и выделить значимые атрибуты
3. Кластеризовать набор данных
4. Разбить набор данных и визуализировать их зависимость
5. Выполнить классификацию данных с дальнейшим обучением модели
6. Выполнить Feature Engineering для разработанной модели
7. 8 апреля построить регрессионную модель на основе обработанных данных
8. Выполнить оценку качества регрессионной модели и выбрать наилучшую из всех для дальнейшей визуализации её работы
9. Разработать программный продукт – виджет для туристического сайта, который будет выполнять действия по оценке опасности посещение той или иной страны с дальнейшим предсказанием о, возможно, будущих заражениях

Оглавление

[Выполнение обработки 5](#_Toc68855541)

[Кластеризация 7](#_Toc68855542)

[Разбить набор данных и визуализировать их зависимость 8](#_Toc68855543)

[Визуализация данных 9](#_Toc68855544)

[Feature Engineering 10](#_Toc68855545)

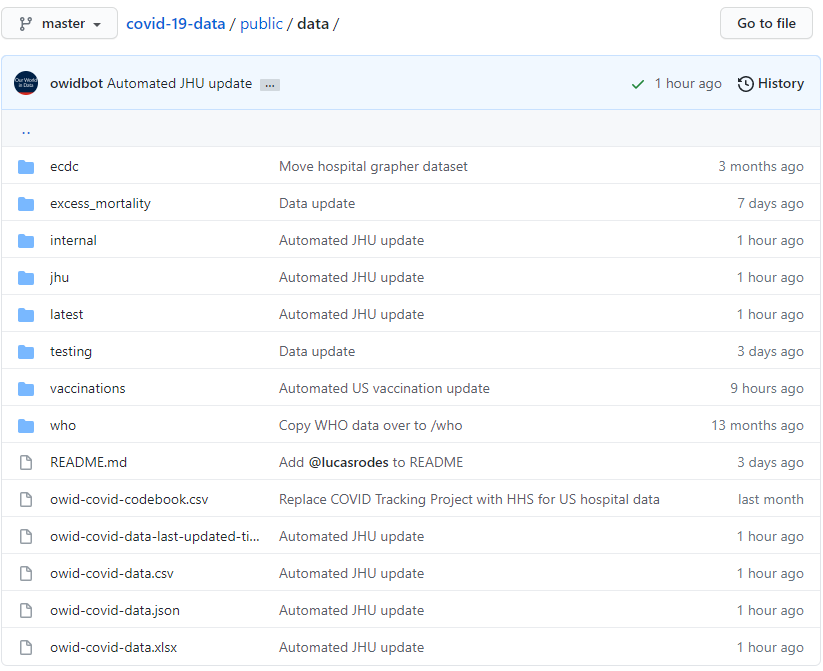
[Построение регрессионной модели на основе обработанных данных 11](#_Toc68855546)

[Разработать программный продукт – виджет для туристического сайта, который будет выполнять действия по оценке опасности посещение той или иной страны с дальнейшим предсказанием о, возможно, будущих заражениях 12](#_Toc68855547)

[Итоговый вариант виджета 13](#_Toc68855548)

1. **Загрузка данных и анализ**

Работу начал выполнять с анализа открытых и предоставленных мне данных.



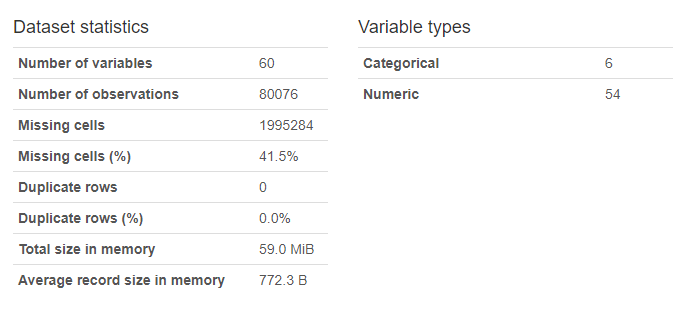
Можно заметить, что некоторые данные загрузили не давно, а значит, что они актуальны для нас.

Делаем парсинг данных и предварительный анализ

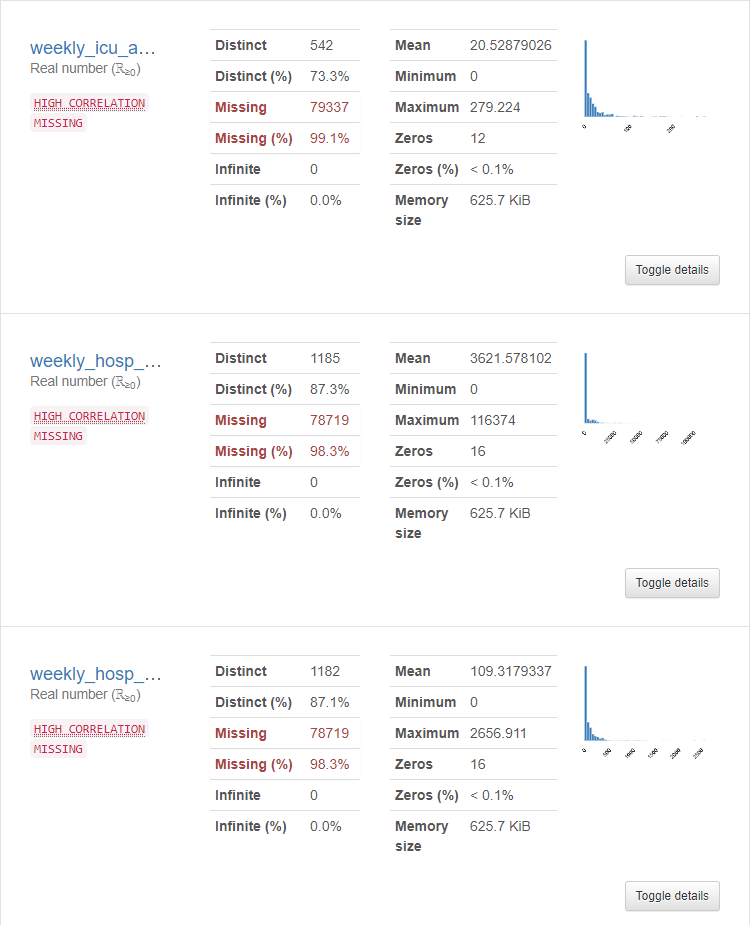
Исходный размер данных, как мы можем видеть, составляет 80295 строк и 58 столбцов.



После анализа датасета библиотекой pandas можем детально посмотреть на данные



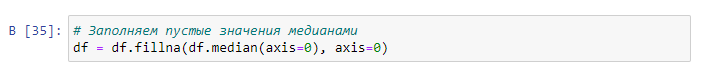
Из анализа видно, что общие пропуски в строках имеют 41.5%, а их общая сумма достигает 1995284.



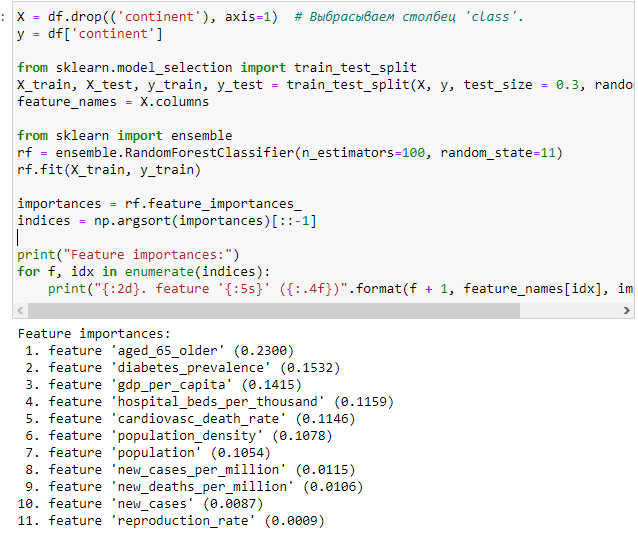
# Выполнение обработки



Удалил ненужные столбцы, которые несли в себе индексы, удалил данные с большим числом пропусков, потому что с ними нельзя было работать из-за их недостоверности. Те данные, в которых процент пропуска был менее 50% заполнил средним значением ближайших соседей.



Выделим наиболее значимые атрибуты при помощи метода rf.feature\_importances



Как мы можем заметить наиболее важными признаками являются:

1. aged\_65\_older
2. diabetes\_prevalence
3. gdp\_per\_capita
4. hospital\_beds\_per\_thousand
5. cardiovasc\_death\_rate

# Кластеризация

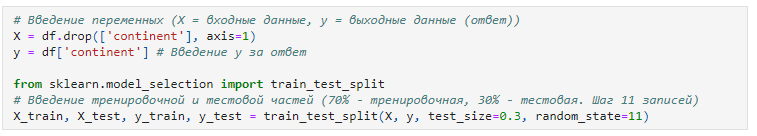
Обработав и подготовив данные для определения уровня опасности посещения страны, нужно было кластеризовать данные, то есть разбить их на группы по схожим признакам для дальнейшего интегрирования их в виджет



Вывод: Из всех методов лучшим оказался метод KMeans - 0.183. На втором месте Spectral - 0.179, на третьем Agglomerative - 0.177, и на четвёртом - Affinity - 0.175. Из результата алгоритмов будем использовать алгоритм KMeans, потому что он показал лучший результат среди остальных.

# Разбить набор данных и визуализировать их зависимость

Следующим шагом для достижения цели по созданию виджета это разбиение данных на обучающую и тестовую выборки для составления модели классификаций



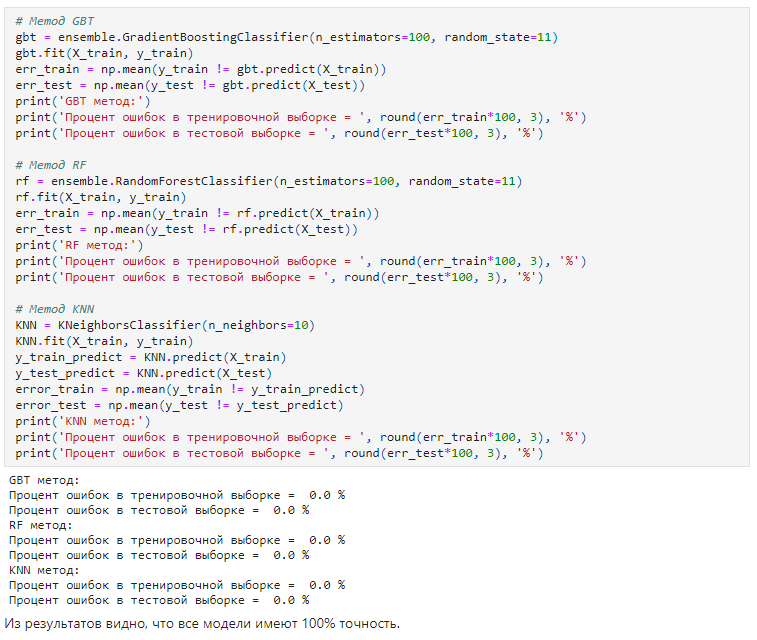
Данные были разбиты сначала на 80% обучающая и 20% тестированная и обучалась эта выборка 4-мя моделями

SVC – модель не смогла обучится и было принято решение не использовать её

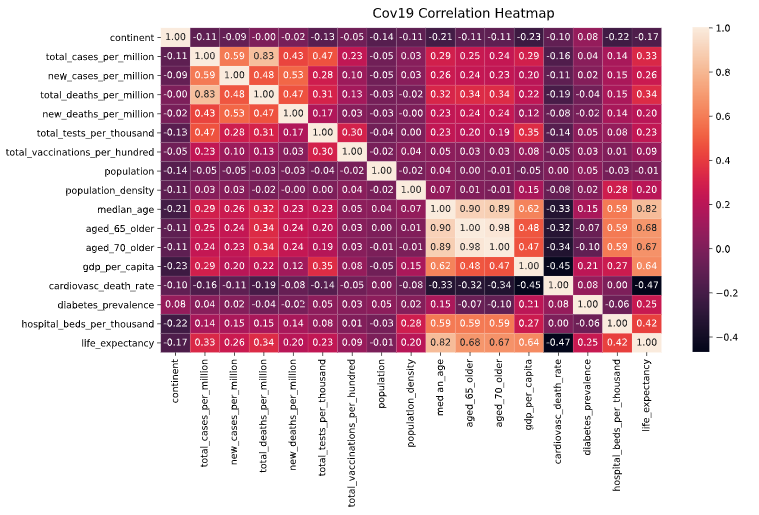
Gradient Boosting Classifier – модель показала 100% точность

Random Forest – модель показала 100% точность

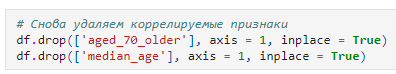
K Neighbors Classifier – модель показала 100% точность

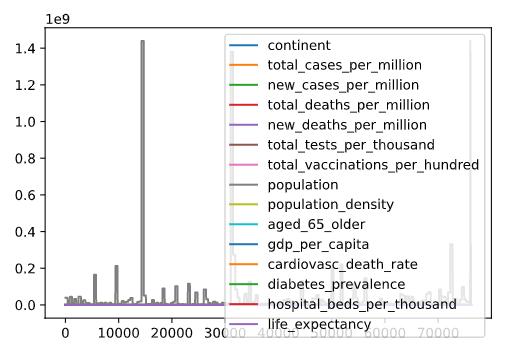


# Визуализация данных



Из визуализации видно, что некоторые данные между собой хорошо коррелируют, а это означает, что признаки имеют одинаковый набор данных между собой, удаляем их для облегчения построения моделей

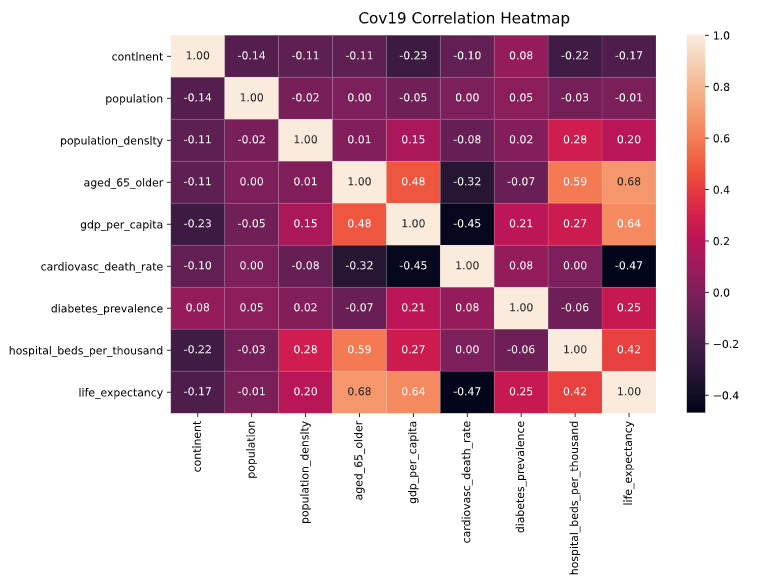




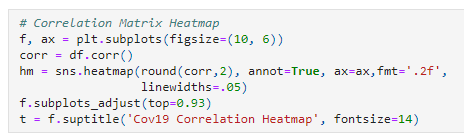
Из следующего графика можно заметить, что данные имеют выбросы данных, которые не являются действительными, тоже обрабатываем их

# Feature Engineering

Оценив свои модели было принято решение попробовать обучить модели на выборке 70% обучающей и 30% тестированной. Такая модель показала тоже 100% результат, и было принято решение оставить эту модель, во избежание её переобучения на тестированных данных

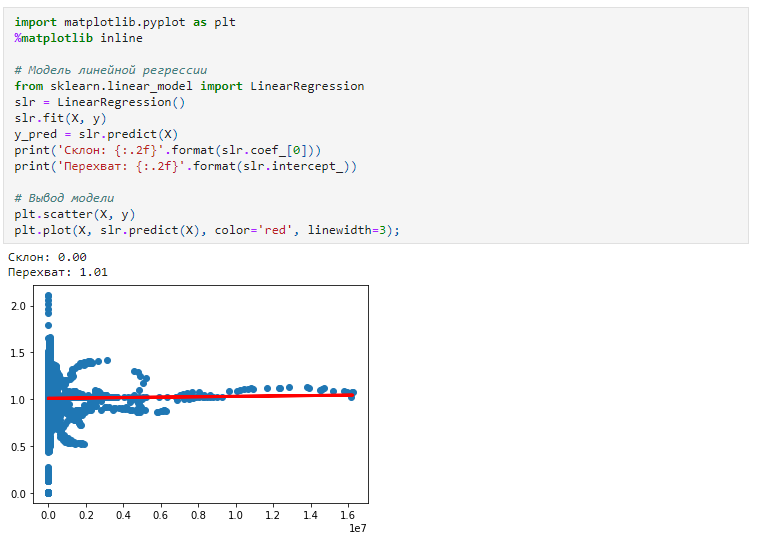


Теперь мы можем наблюдать тепловую карту корреляции без похожих признаков

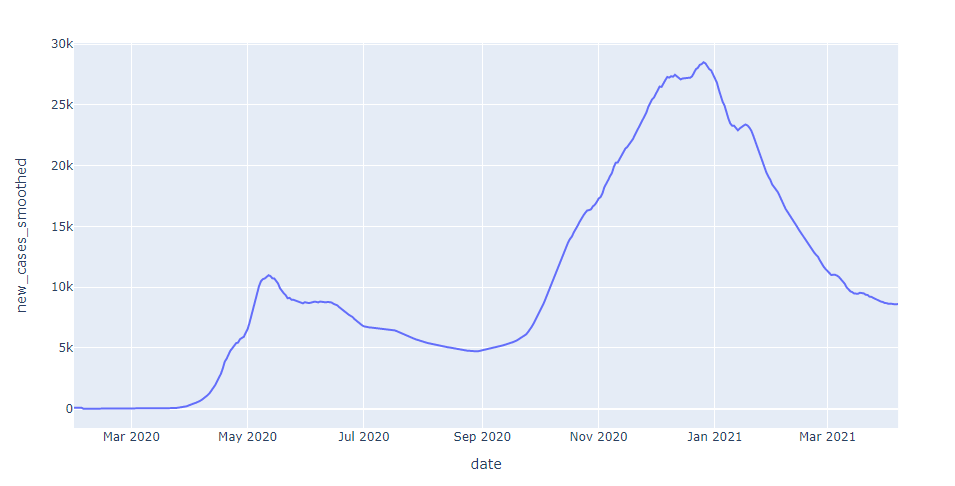


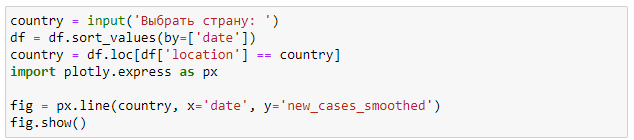
# Построение регрессионной модели на основе обработанных данных

Для дальнейшего продвижения к достижению цели создать интерактивный виджет – помощник нужно построить регрессионную модель на основе обработанных данных за первый день



Как можно заметить, линейная регрессия с одной независимой переменной показывает неудачные результаты, так как является практически параллельной оси абсцисс. Поэтому предсказание будет одним и тем же. Но модель несмотря на это получилась довольно точной





# Разработать программный продукт – виджет для туристического сайта, который будет выполнять действия по оценке опасности посещение той или иной страны с дальнейшим предсказанием о, возможно, будущих заражениях

Виджет должен быть не только технологичным и функциональным, но ещё и лаконичным с приятным дизайном, который мог бы подстраиваться под дизайн любого сайта.



Обозначение маркеров  
1. Ввод названия страны;  
2. Название графика;  
3. Выбор промежутка времени;  
4. Светофор;  
5. Опасное время;  
6. Новые случаи заболеваний;  
7. Дата.

Это самый первый макет, разработанный на скорую, ведь надо было делать отчёт для заказчика.

# Итоговый вариант виджета



Итоговый вариант виджета представлен на скриншоте. Пользователь может удобным и понятным способом выбрать интересующую его страну или вписать её самостоятельно в соответствующее поле, выбрать интересующий промежуток посещения той или иной страны от 1 дня до полугода и увидеть рекомендуемое состояние по поездке в страну.

Для разработчиков тоже есть приятные вещи. Виджет легко встраивается на любую HTML страничку с настройками под любой дизайн сайта (от изменения цветовой палитры графика, до масштабирования отдельных объектов таких как: выбор страны, промежутка и уровня опасности). Мой виджет поможет не только туристам, но и государственным сайтам, которые принимают участие в борьбе против COVID-19