**Федеральное государственное бюджетное образовательное учреждение высшего образования**

**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ**

**РОССИЙСКОЙ ФЕДЕРАЦИИ»**

**Департамент анализа данных,**

**принятия решений и финансовых технологий**

**Отчет по контрольной работе**

По дисциплине

**Технологии обработки больших данных**

**По теме**

**«****Анализ эпидемиологических и экономических показателей в период пандемии коронавируса»**

Выполнил:

Студент группы ПИ17-1

Носко Максим Михайлович

Преподаватель:

Корчагин Сергей Алексеевич

Москва

2020

Содержание

[Введение 2](#_Toc59743643)

[Получение и обработка данных 3](#_Toc59743644)

[Применение параллельных вычислений 7](#_Toc59743645)

[Анализ данных 9](#_Toc59743646)

[Хранение данных 19](#_Toc59743647)

[Вывод 20](#_Toc59743648)

# Введение

На сегодняшний день в мире зафиксировано более 54 миллионов случаев заражения коронавирусом. Это число растёт увеличивающимися темпами и скоро количество случаев заражения превысит 1% населения Земли. В России уже зарегистрировано почти 2 миллиона случаев заражения, то есть почти 1,5% населения России. Пандемия оказала большое влияние на жизнь как россиян, так и людей во всех странах мира, а значит весьма вероятно сходство динамики экономических показателей и эпидемиологических.

Для анализа данных больших объёмов применялись следующие технологии:

* Python – высокоуровневый язык программирования общего назначения, ориентированный на повышение производительности разработчика и читаемости кода.
* Anaconda – дистрибутив языков программирования Python и R, включающий набор популярных свободных библиотек, объединённых проблематиками науки о данных и машинного обучения.
* Pandas – программная библиотека на языке Python для обработки и анализа данных. Работа pandas с данными строится поверх библиотеки NumPy, являющейся инструментом более низкого уровня.
* Matplotlib – библиотека на языке программирования Python для визуализации данных двумерной графикой.
* Parquet – бинарный, колоночно-ориентированный формат хранения больших данных.

Для работы с данными применялась параллельная обработка больших данных, которая была реализована с помощью потоков. Поток выполнения – наименьшая единица обработки, исполнение которой может быть назначено ядром операционной системы. Реализация потоков выполнения и процессов в разных операционных системах отличается друг от друга, но в большинстве случаев поток выполнения находится внутри процесса. Несколько потоков выполнения могут существовать в рамках одного и того же процесса и совместно использовать ресурсы, такие как память, тогда как процессы не разделяют этих ресурсов.

# Получение и обработка данных

Данные по динамике распространения коронавируса были взяты из открытого репозитория Университет Джонса Хопкинса в GitHub.

Так как данные публикуются в виде csv файлов, для взаимодействия с ними была выбрана библиотека Pandas (рис. 1).



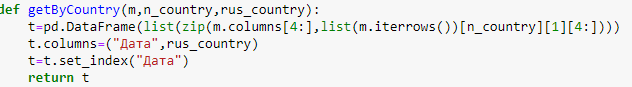
*Рис. 1: чтение csv-файла с помощью библиотеки Panda*s

Исходная таблица представляла из себя количество подтверждённых случаев заражения коронавирусом по странам (строки) и по дням (столбцы) (рис. 2).

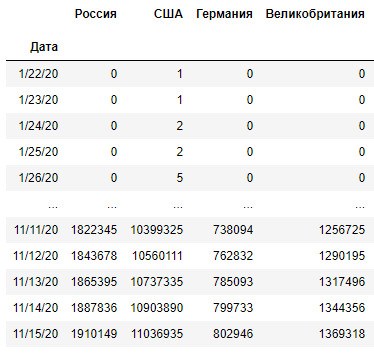


*Рис.2: исходная таблица с данными*

Для начала, из исходной таблицы были выделены несколько стран, и была сформирована, с помощью функции, позволяющей извлекать из исходных данных данные по стране (рис. 3) новая таблица, в которой страны были столбцами, а даты – заголовками строк, что позволило в дальнейшем удобнее производить сравнительный анализ рассматриваемых данных (рис. 4).

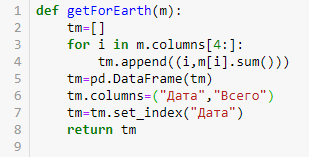


*Рис. 3: функция, извлекающая данные по стране из исходных данных*



*Рис. 4: преобразованная таблица*

Для дальнейшего анализа также понадобится общее число заболевших по всему миру, для чего также была написана функция (рис. 5), и в таблицу был добавлен соответствующий столбец (рис. 6).



*Рис. 5: функция, суммирующая число заболевших по дате*

**

*Рис. 6: общее число заболевших*

Также для дальнейшего анализа потребуется сравнивать не только абсолютное количество случаев заболевания, но и долю заражённых из населения страны, так как именно доля заболевших от всех людей в стране имеет ключевое влияние на экономику страны. На основе данных последних переписей населения в соответствующих странах, было произведено вычисление процентов с помощью функции (рис. 7), и в таблицу были добавлены соответствующие столбцы (рис. 8).

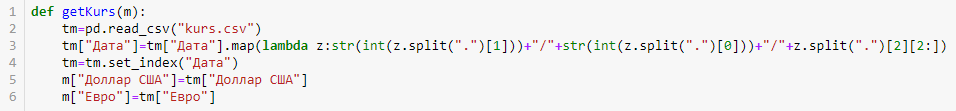
**

*Рис. 7: функция, добавляющая в таблицу процент заболевших коронавирусом из всего населения страны*

**

*Рис. 8: столбцы с долей заболевших в процентах*

Для анализа динамики экономической ситуации было решено рассматривать колебания курсов трёх валют: рубля, доллара США и евро. Данные по курсу были взяты с сайта Центрального банка Российской Федерации и преобразованы в csv-формат, после чего была написана функция (рис. 9), добавляющая в таблицу соответствующие столбцы (рис. 10).

**

*Рис. 9: функция получения курсов валют*

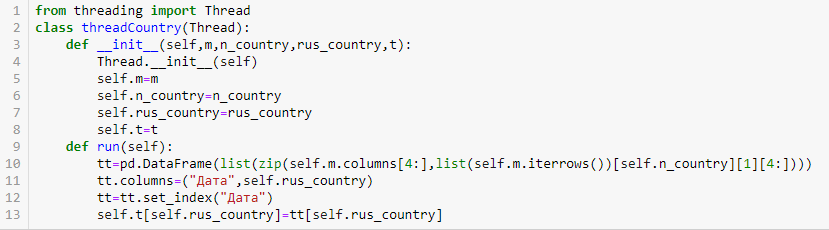
**

*Рис. 10: столбцы со значениями курсов в определённые даты*

# Применение параллельных вычислений

В обработке данных присутствует этап, который представляет из себя несколько одинаковых и независимых задач, что означает, что данные операции можно выполнять параллельно, например, в различных потоках.

Для приведения к параллельному виду, задача отбора информации по стране была преобразована к классу, который наследуется от класса потока – Thread (рис. 11), после чего были сделаны измерения быстродействия программы без использования параллельных вычислений (рис. 12) и с их использованием (рис. 13).

**

*Рис. 11: класс потока параллельной обработки данных по стране*

**

*Рис. 12: последовательная обработка данных*

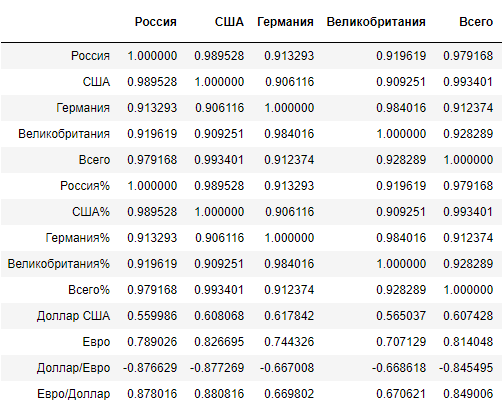
**

*Рис. 13: эффективность использования параллельных вычислений*

Несмотря на то, что общее время обработки данных уменьшилось только на 2%, обработка данных по странам ускорилась на 7%, что весьма значимо в обработке больших данных.

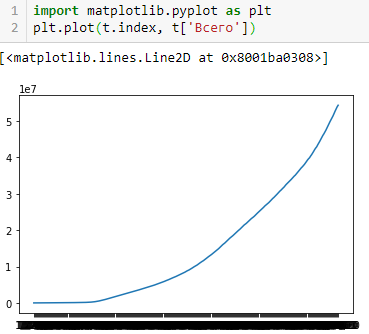
# Анализ данных

В первую очередь, была рассмотрена матрица корреляции (рис. 14) из которой видно, что динамика заражений во всех странах, в целом, очень похожа, так как значения коэффициентов корреляции не опускаются ниже 0,9. При этом наиболее высокая корреляция наблюдается у заражений в США и заражений во всём мире, что логично, так как в США количество случаев заражений наибольшее. При этом, динамика заражений в России и США, например, весьма похожа, и корреляция между ними составляет 0,99, в то время как корреляция динамики распространения вируса между США и Германией составляет 0,91.

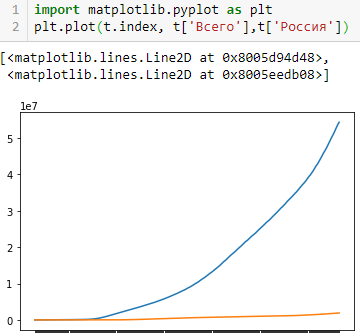
**

*Рис. 14: корреляция динамики распространения коронавируса*

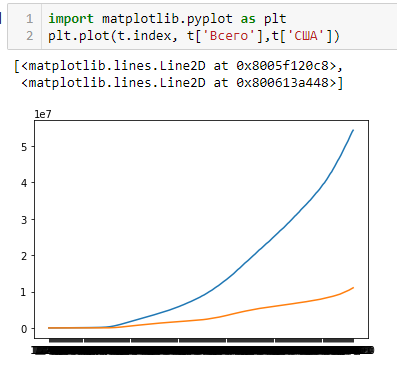
Рассмотрим общую динамику распространения вируса во всём мире – был составлен график общей динамики (рис. 15), и на него были добавлены для сравнения графики распространения вируса в России (рис. 16), в США (рис. 17) и в Германии (рис. 18).

**

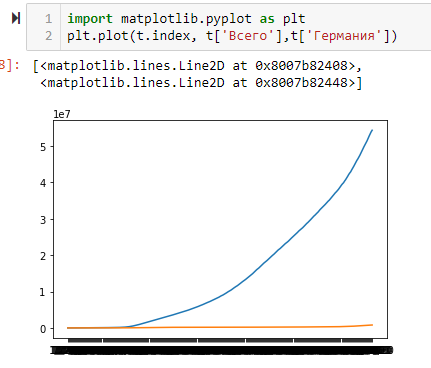
*Рис. 15: график общемировой динамики распространения вируса*

**

*Рис. 16: график распространения вируса во всём мире (синий) и в России (рыжий)*

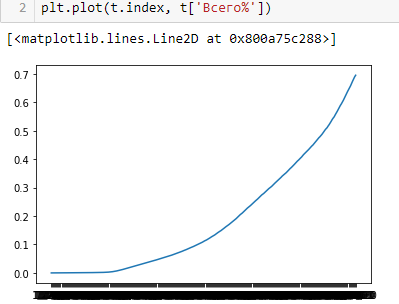
**

*Рис. 17: график распространения вируса во всём мире (синий) и в США (рыжий)*

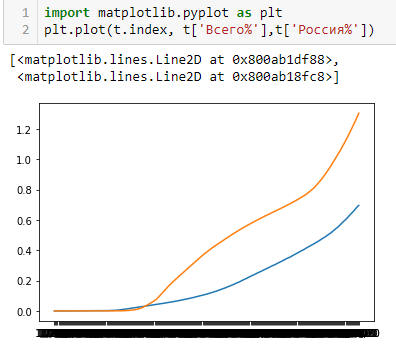
**

*Рис. 18: график распространения вируса во всём мире (синий) и в Германии (рыжий)*

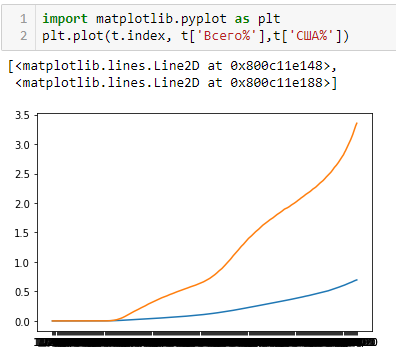
Как видно из графиков, хотя корреляция у динамики распространения вируса в России и в США весьма большая, в мировом масштабе ситуация в России больше схожа с ситуацией в Германии. Однако это абсолютные значения, поэтому для более взвешенного сравнения ситуаций были составлены графики распространения вируса по отношению к количеству людей в стране: был составлен график общей динамики (рис. 19), и на него были добавлены для сравнения графики распространения вируса в России (рис. 20), в США (рис. 21) и в Германии (рис. 22).



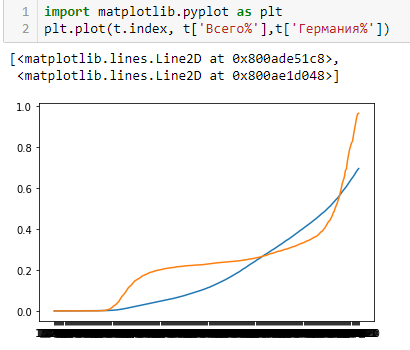
*Рис. 19: график общемировой динамики распространения вируса*



*Рис. 20: график распространения вируса во всём мире (синий) и в России (рыжий)*



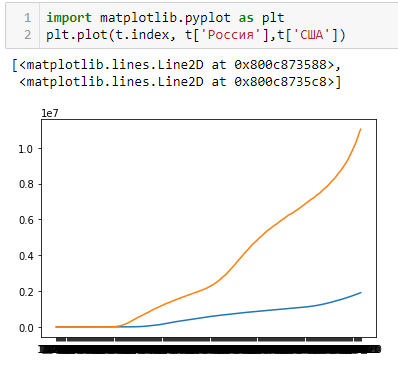
*Рис. 21: график распространения вируса во всём мире (синий) и в США (рыжий)*



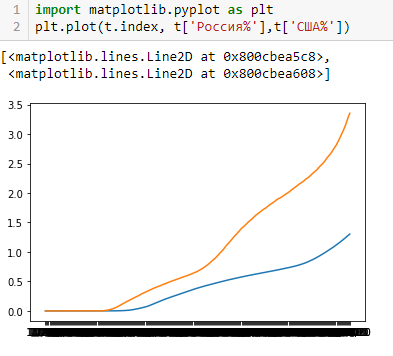
*Рис. 22: график распространения вируса во всём мире (синий) и в Германии (рыжий)*

Как видно из графиков, в сравнении с мировым распространением, распространение вируса в России и в США происходит весьма быстрее, чем в среднем на планете. При этом ситуация в Германии примерно похожа на всемирную, и довольно длительный период даже была меньше, чем в среднем во всём мире.

В США отмечается значительно большее число заболевших, чем в России, но чтобы рассмотреть скорость распространения вируса относительно населения, необходимо было составить не только абсолютный сравнительный график (рис. 23), но и процентный (рис.24).



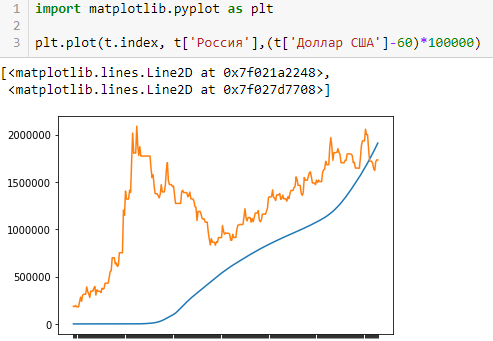
*Рис. 23: график распространения вируса по количеству человек в России (синий) и в США(рыжий)*



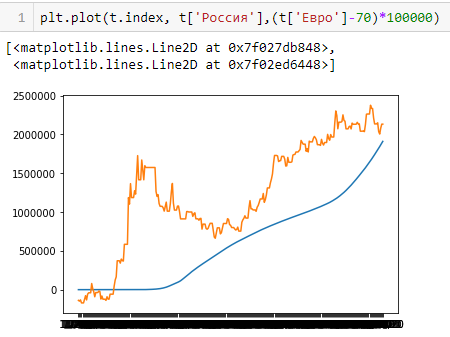
*Рис. 24: график распространения вируса по количеству человек относительно населения страны в России (синий) и в США(рыжий)*

Как видно из графиков, в процентном соотношении вирус распространяется в США также весьма быстрее, чем в России, но разница между ситуациями не настолько различается.

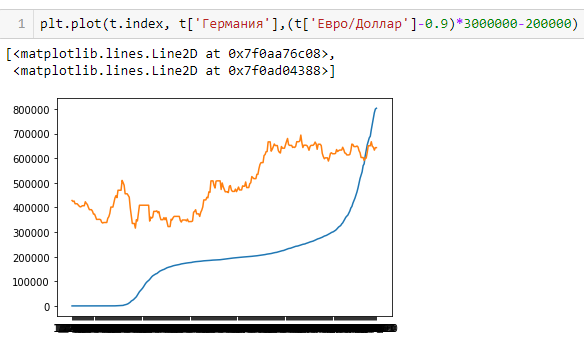
Для анализа корреляции экономической ситуации были составлены графики распространения вируса в России в сравнении с динамикой курса доллара (рис. 25), а также курса евро (рис. 26), а также аналогичные графики, сравнивающие динамику распространения вируса в Германии (рис. 27) и в США (рис. 28) с сравнении с динамикой курса евро и доллара. При этом, так как порядок чисел значений и дисперсии у данных показателей разный, было произведено математическое преобразование экономических показателей.



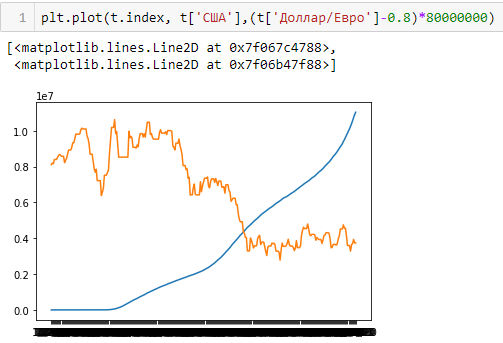
*Рис. 25: график распространения вируса по количеству человек в России (синий) и график динамики курса доллара США(рыжий)*



*Рис. 26: график распространения вируса по количеству в России (синий) и график динамики курса евро(рыжий)*



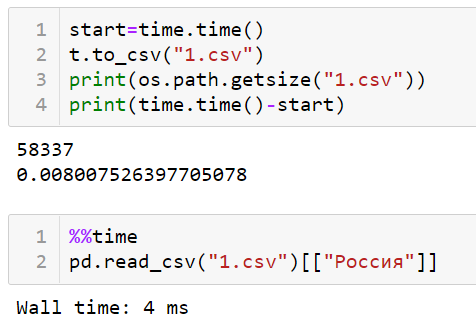
*Рис. 27: график распространения вируса по количеству человек в Германии (синий) и график динамики* *соотношения евро к доллару(рыжий)*



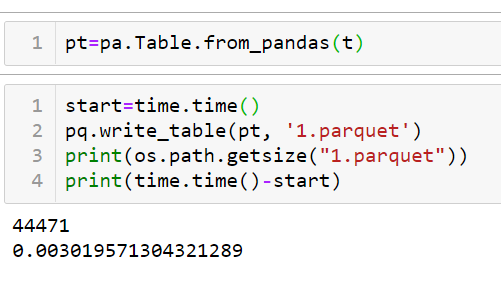
*Рис. 28: график распространения вируса по количеству человек в США (синий) и график динамики соотношения доллара к евро(рыжий)*

# Хранение данных

Для хранения данных, в особенности, данных больших объёмов, можно использовать как обычный, текстовый и читаемый человеком формат csv (рис. 29), так и, например, двоичный формат parquet (рис. 30).

**

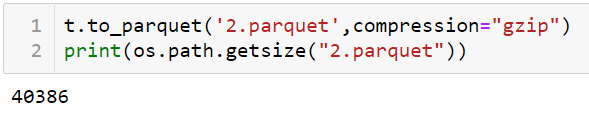
*Рис. 29: сохранение данных в формате csv*



*Рис. 30: сохранение данных в формате parquet*

При сохранении данных в формате parquet получилось сократить занимаемое данными место более чем на 31%, что является весьма хорошим показателем эффективности.

При этом, ещё большей эффективности можно добиться, если при записи в формате parquet использовать сжатие (рис. 31). В таком случае, итоговый выигрыш в занимаемом на диске месте будет составлять более 44%.

**

*Рис. 31: сохранение данных в формате parquet с использованием сжатия*

# Вывод

Анализ эпидемиологических и экономических данных показал, что обострение ситуации с распространением коронавируса в России и США совпадает с уменьшением ценности национальных валют относительно евро, в то время как распространение вируса в Германии не так сильно сказалось на национальной экономике.

Для обработки данных параллельные вычисления показали свою большую эффективность, из-за чего их использование выгодно для обработки больших данных.

Для хранения больших данных эффективным решением является их сохранение в сжатом виде в двоичном формате, что позволяет экономить значительные объёмы дискового пространства.