2022

Анализ данных по коронавирусу

ПОИСК ИНТЕРЕСНЫХ ЗАКОНОМЕРНОСТЕЙ И ВЗАИМОСВЯЗЕЙ

ШАМСУТДИНОВ АЯЗ АСХАТОВИЧ

Оглавление

1. Введение

Уже с первого курса, когда я только поступил, я хотел заниматься дополнительно вне стен университета. Причин этому было несколько, но основная, наверное, это желание изучить что-то новое и проводить свободное время с пользой. Но на первом курсе у меня было мало знаний, чтобы попасть куда-либо, однако я нашел альтернативу в виде онлайн-курсов.

Так как на моем направление сначала изучали R, я тоже решил больше времени посвятить ему и прошел курс по основам программирования на R(https://stepik.org/course/497/syllabus), который показался мне очень интересным. Дальше в рекомендациях я нашел курс по анализу данных в R(https://stepik.org/course/129/), который я впоследствии и взял за основу при выполнении данной работы. Этот курс был одновременно и легким, и непонятным. Его легкость заключалась в простоте заданий, которые выполнялись по алгоритму из видеоуроков, а сложность в том, что понимание видео требовало знаний математической статистики, которых у меня не было. Где-то на половине я забросил его с мыслью продолжить в будущем, что и случилось.

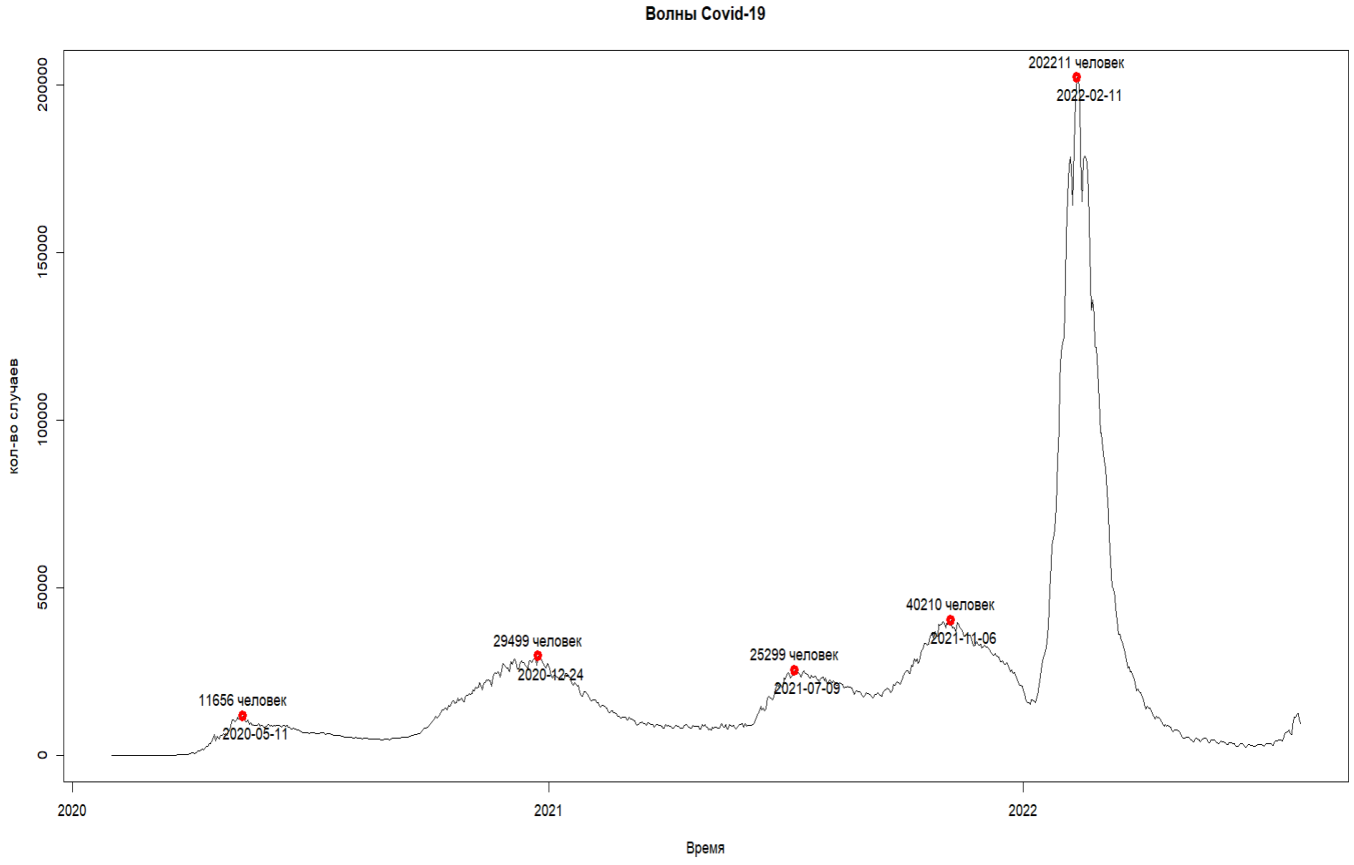
Также упомяну про Python. Его я начал изучать после курса по анализу данных в R и тоже на платформе stepik. Так как у меня уже были навыки программирования в R, который очень похож на python, я довольно быстро прошел один из вводных курсов(https://stepik.org/course/67/syllabus). Продолжить я решил курсом по алгоритмам(https://stepik.org/course/217/syllabus), так как именно он сочетал в себе возможности программировать и применять некоторые знания из математики, которую, к слову, я очень любил. На этом закончился для меня первый год обучения в университете.

Что же касается самой работы, то я, как уже упоминал ранее, пользовался знаниями из курса по анализу данных в R, но также изучал статьи с таких ресурсов как https://www.rdocumentation.org/, https://pozdniakov.github.io/tidy_stats/, https://r-analytics.blogspot.com/ и в целом пытался за короткий промежуток времени изучить основы математической статистике и перенести эти знания в R. Насколько хорошо получилось, покажет данная работа.

2. Первоначальное исследование

Как только я сел выполнять задание, мне в голову пришла идея проверить насколько сильно количество новых заболевших влияет на количество новых вакцинаций. Я подумал, что при появлении новой волны коронавируса, количество новых вакцинаций будет также расти. Что-ж, давайте проверим.

Первым я решил найти количество волн в России:

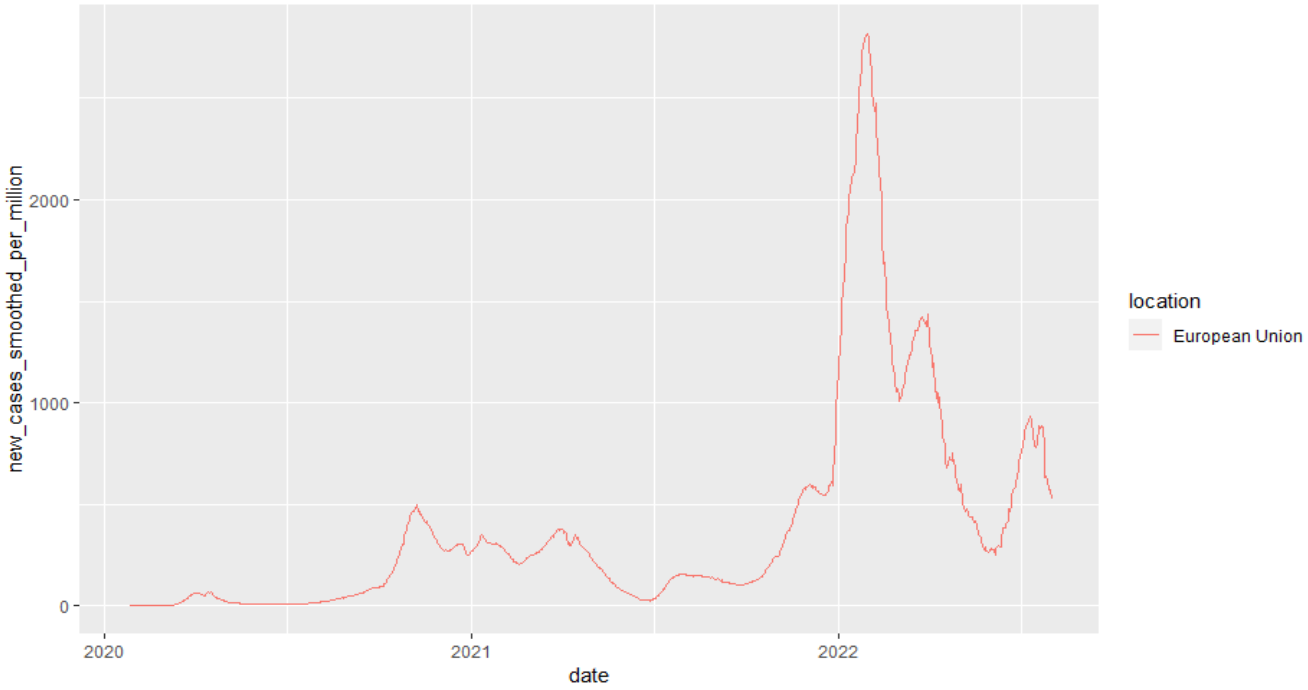


Их оказалось пять: пик 1 волны – 11 мая 2020 года, пик 2 волны – 24 декабря 2020 года, пик 3 волны – 9 июля 2021 года, пик 4 волны – 6 ноября 2021 года и пик 5 волны – 11 февраля 2022 года.

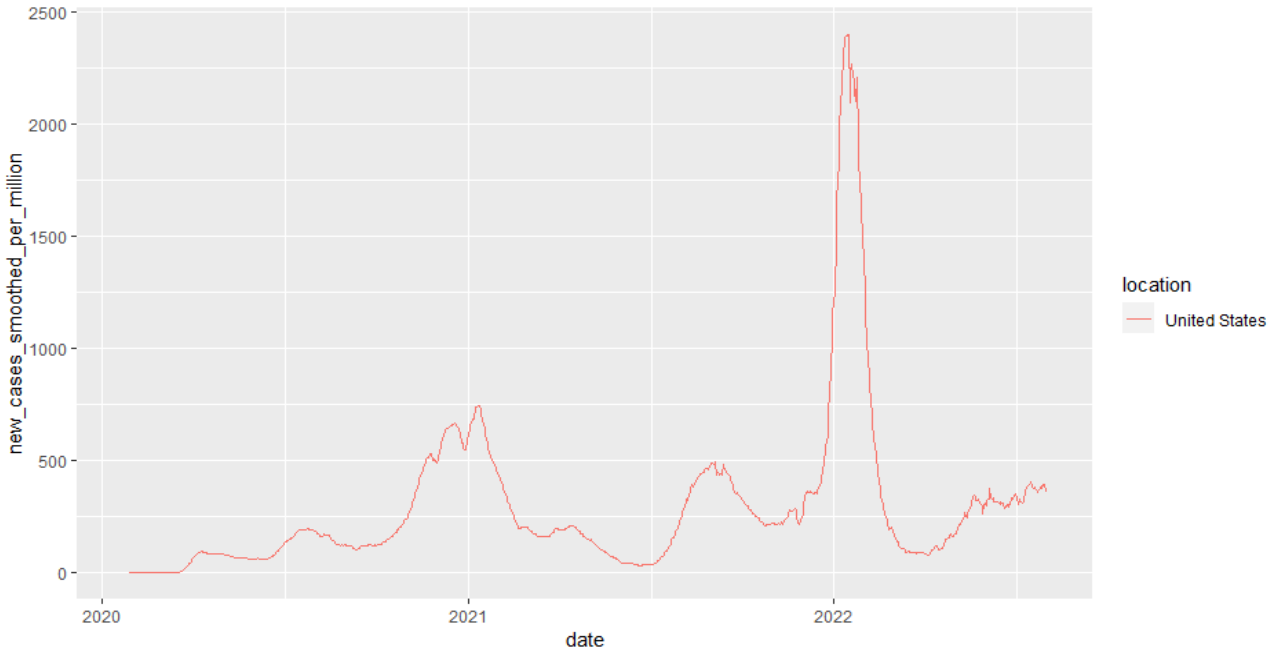
Из графика можно увидеть, что первые четыре волны примерно одинаковые по размеру и среднее количество заболевших было в районе 27 тысяч человек. Однако пятая волна совершенна непохожа на четыре оставшиеся и максимальная точка достигала отметки в 202 211 человек, что примерно в 7,5 раз больше других волн. Странность этим данным добавляет и то, что уже к 2022 году прививка существовала почти как год, а в 2020 году ее даже не было.

Порывшись в интернете, я узнал, что причиной такому большому количеству заболевших в день является как и новый омикрон-штамм, так и новые заболевания даже тех, кто уже привит. Так как оба этих явления влияли на весь мир, я решил заодно посмотреть на данные по заболевшим в других странах:

Европейский союз



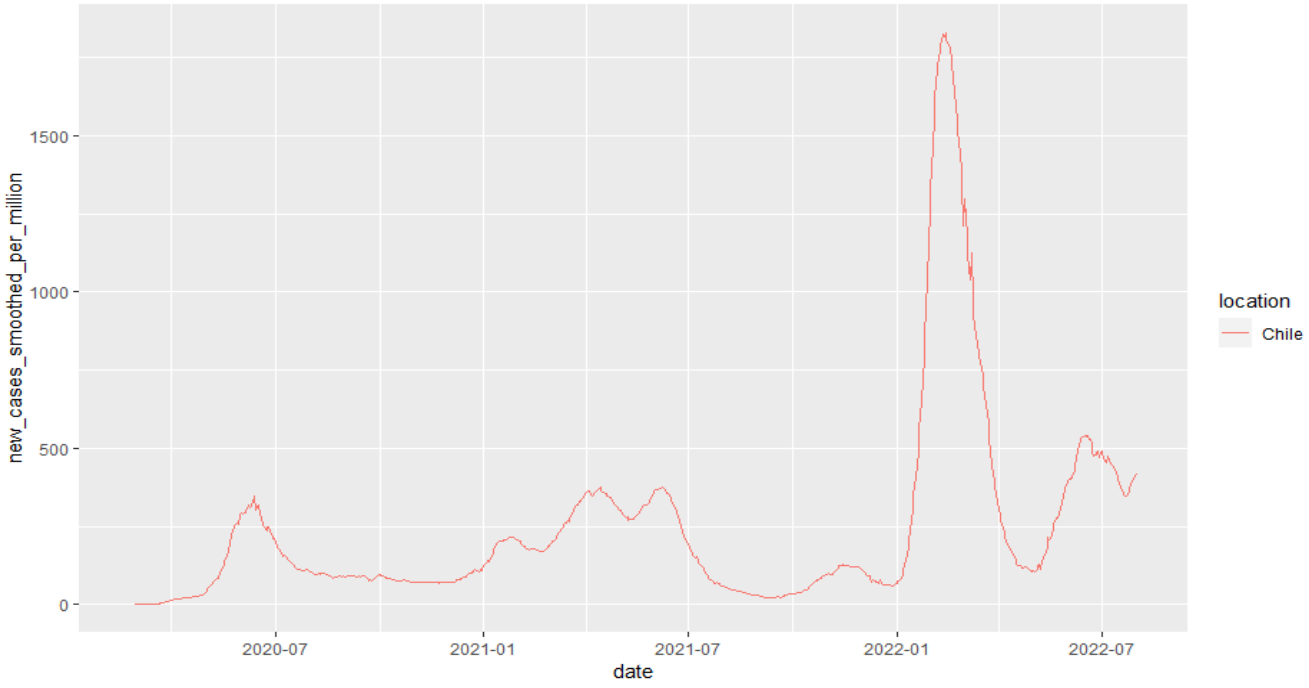
Соединенные Штаты



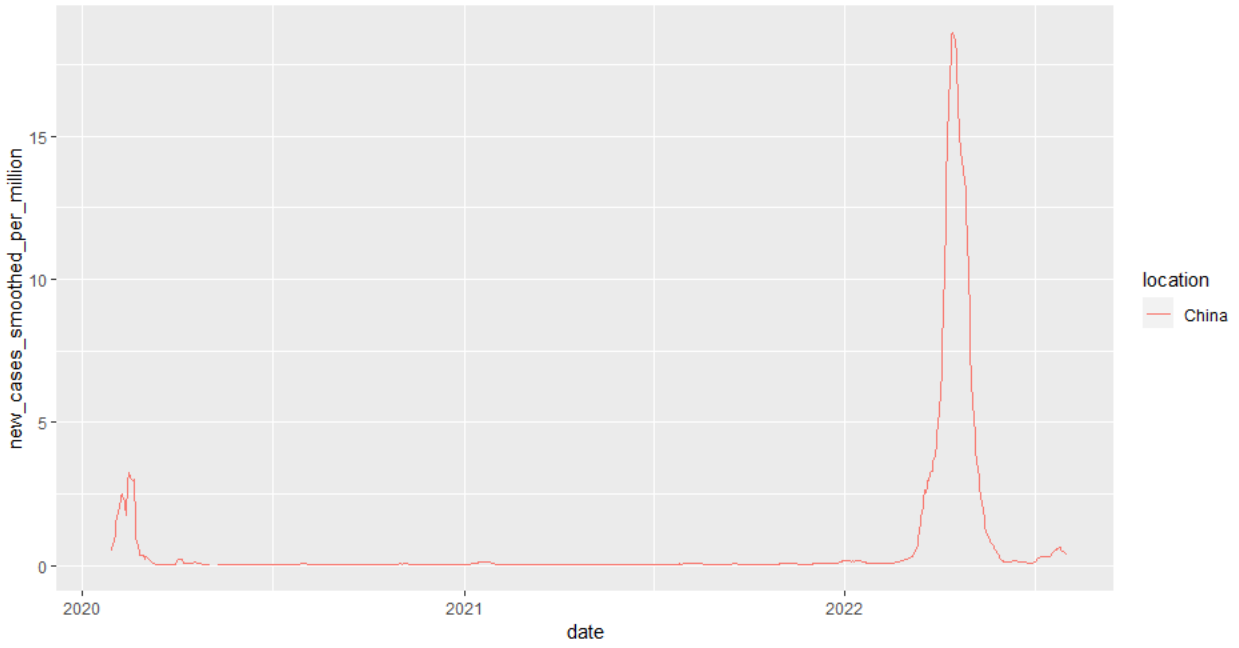
Из этих графиков видно, что в других местах также наблюдается данная тенденция.

Далее я взял страны, в которых очень высокий уровень вакцинированных, чтобы убедиться, что доля вакцинированных на самом деле не уменьшает количество заболеваемых:

Чили (96% вакцинированных)

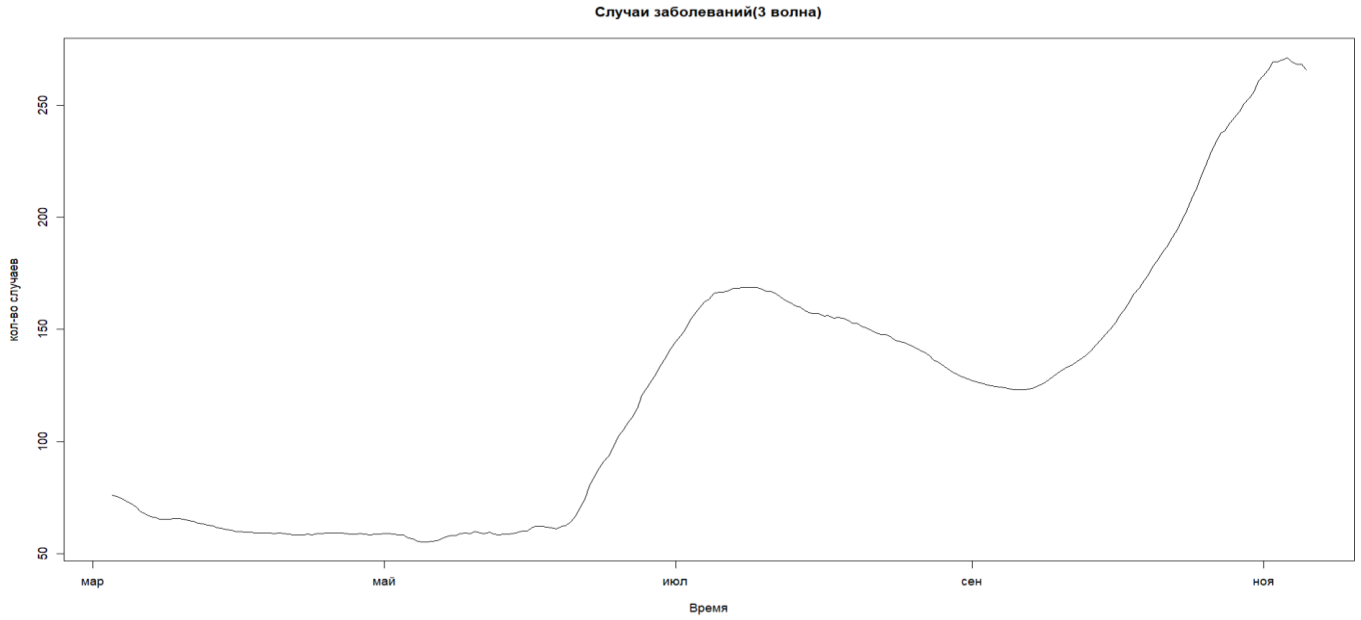


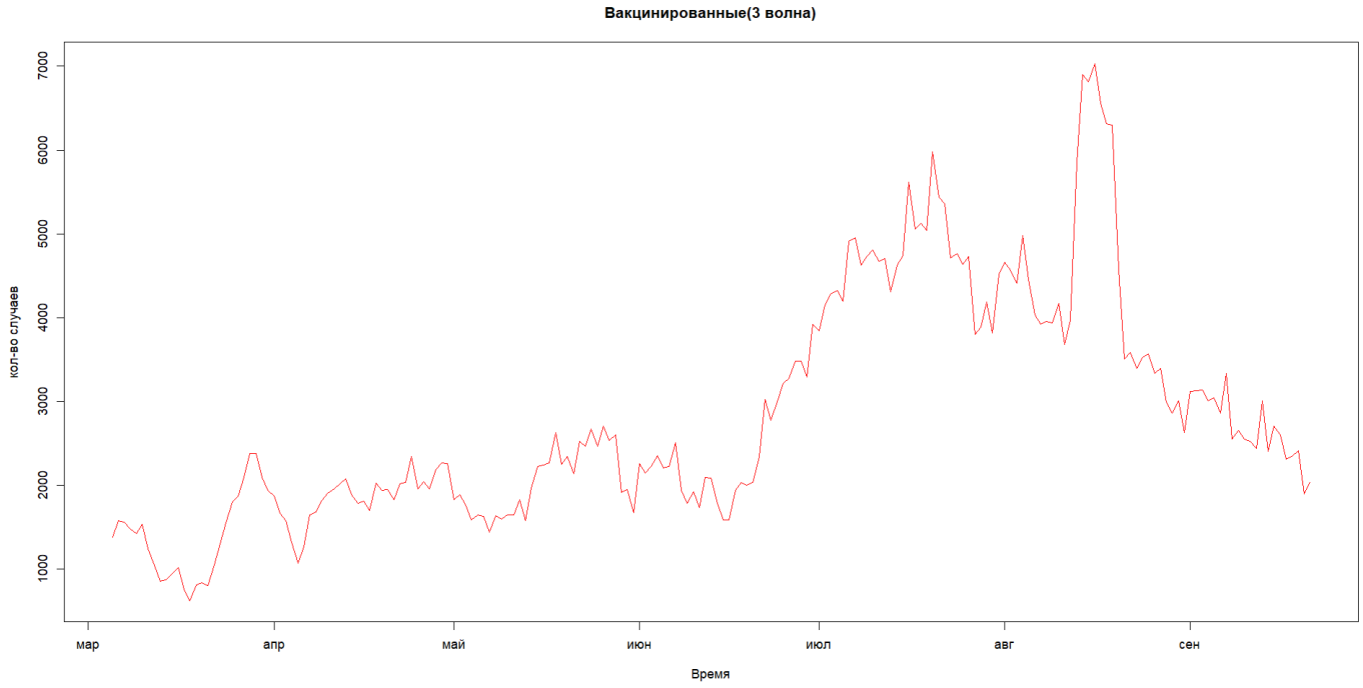
Китай (92% вакцинированных)



Таким образом, во многих странах пик заболевших в 2022 году во много раз превосходит пики заболеваний в другие периоды и не зависит от процента вакцинированных.

Но я немного отвлекся от темы. Для того, чтобы понять, играет ли роль новая волна вируса в увеличении вакцинированных, я сравню графики новых вакцинированных и новых случаях по волнам. Массовую вакцинацию стали проводить лишь с 2021 года, поэтому первые две волны анализировать не получится. Начнем с 3-ей волны:



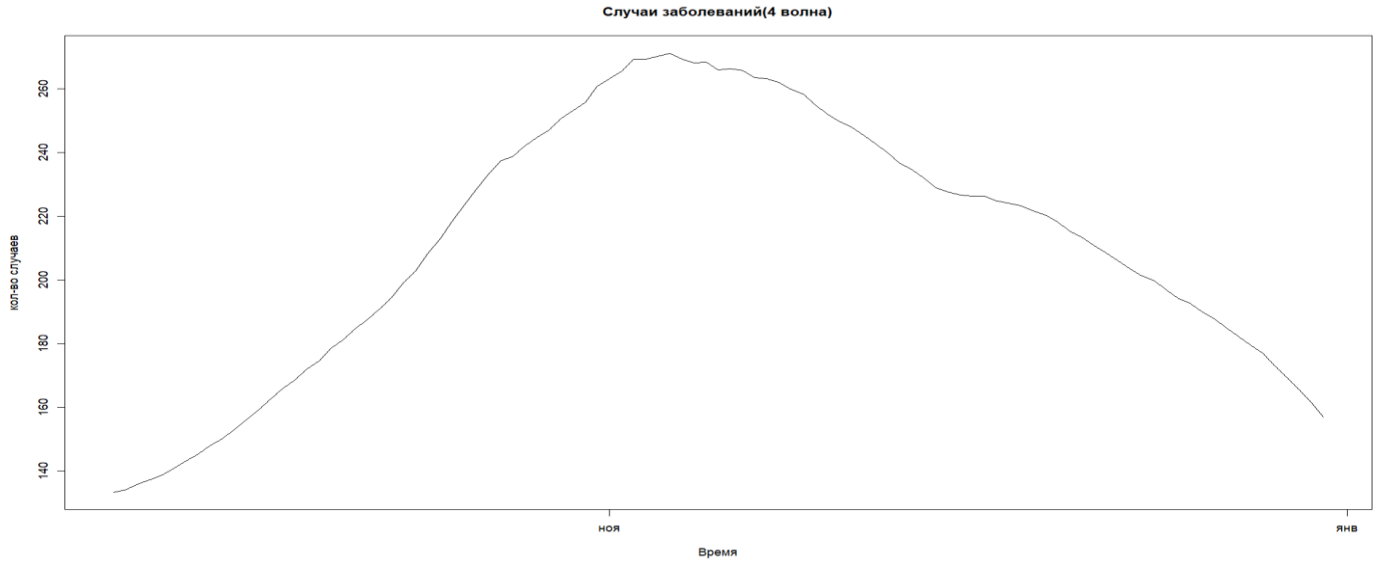


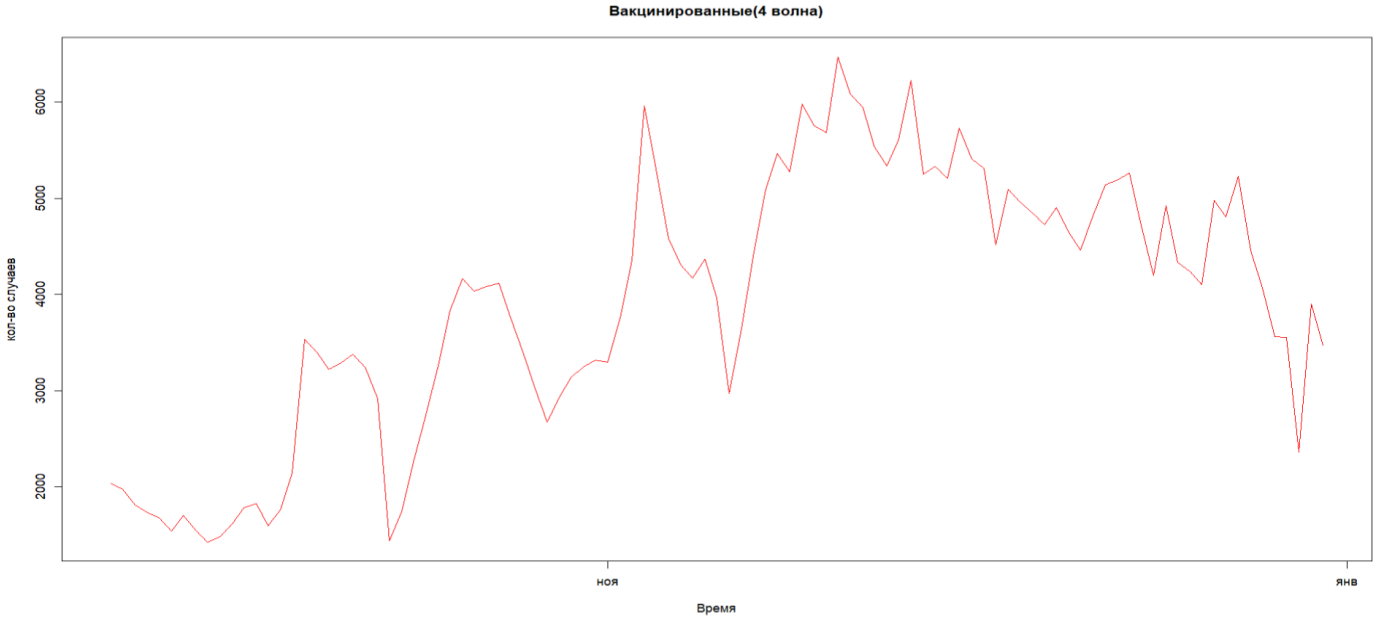
Если внимательно присмотреться к графикам, то можно увидеть, что количество заболевших начало расти в начале июня и примерно к началу июля взошло на пик, когда количество вакцинаций начало расти в середине июня, к июлю достигло половины пика и лишь к середине июля достигло своего локального экстремума. Однако, если количество заболеваний начало падать до сентября, то количество

вакцинаций в середине августа достигло еще одного экстремума, который больше июльских пиков, что привносит неясность в эти данные.

Соответственно, лаг между появлением новой волны и ростом вакцинаций составил примерно полмесяца. Также важно заметить, что в обоих случаях время, которое было необходимо для достижения пика составило примерно месяц, однако количество заболевших, достигнув максимума, пошло на спад, а количество вакцинированных через время снова пошло обновлять пики.

Теперь рассмотрим 4-ую волну:



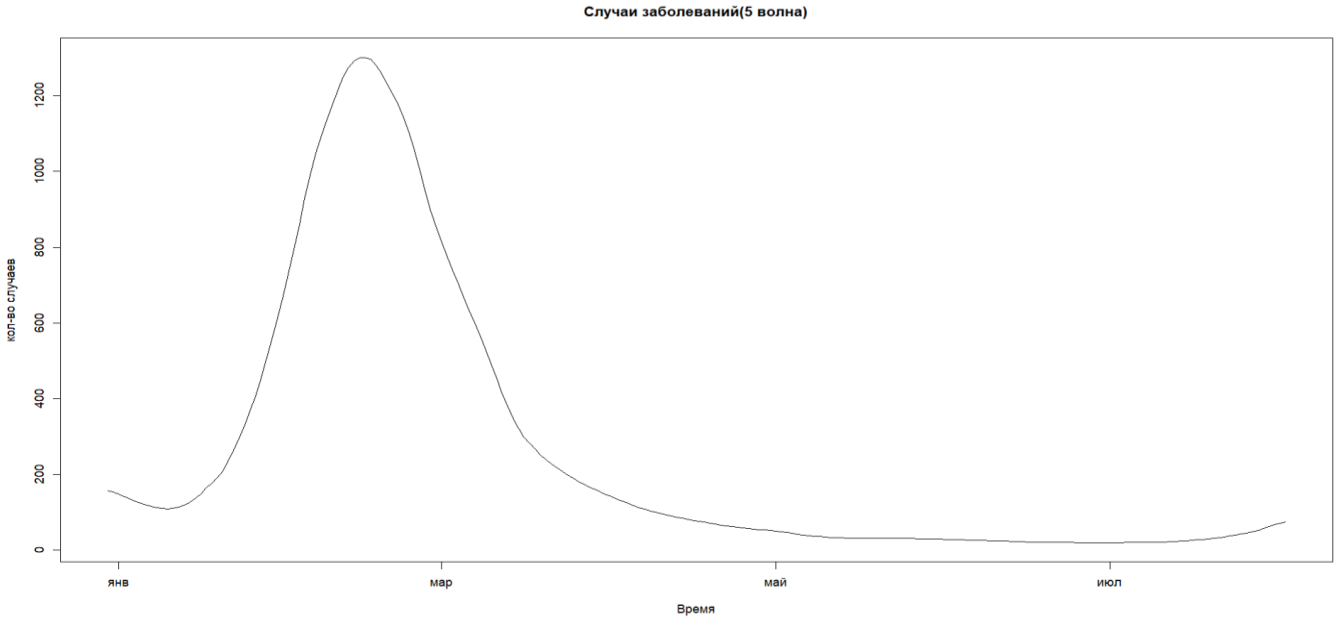


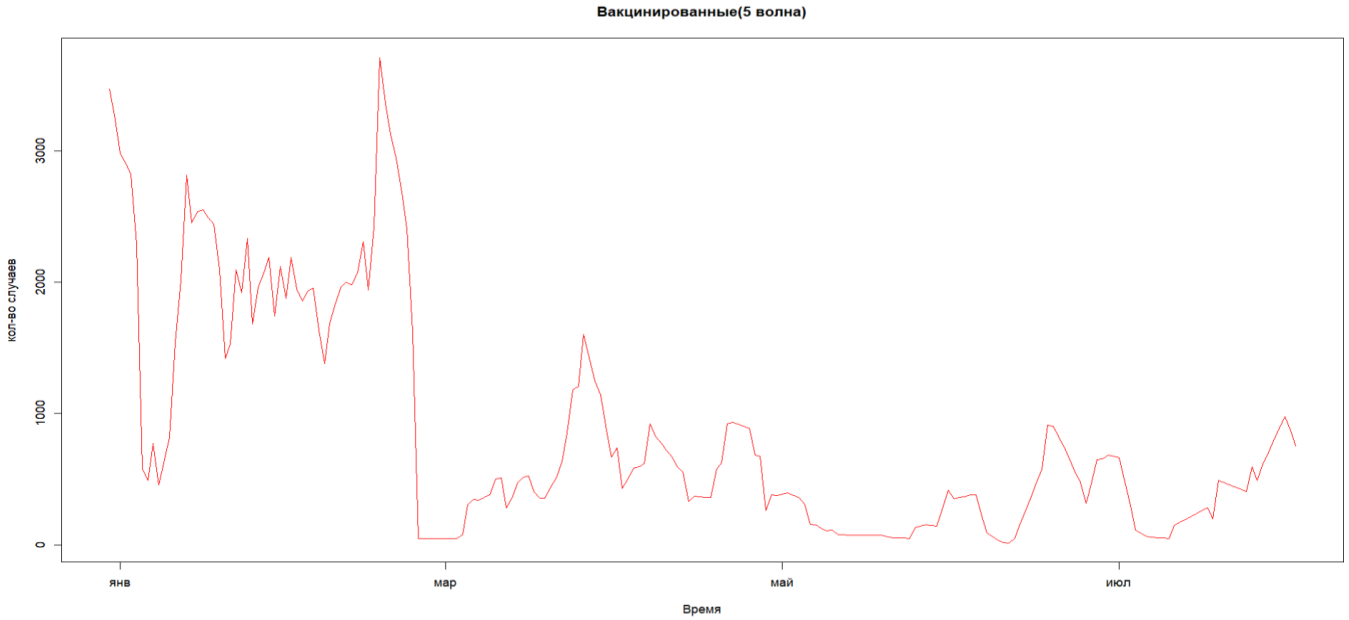
Можно увидеть, что график заболеваний имеет вид прямой в то время как график вакцинированных больше похож на ломаную с большим количеством локальных минимумов. Смею предположить, что это связано с неверными данными, из-за чего и можно видеть такие большие провалы. Несмотря на эти провалы, видно, что в обоих графиках идет постепенное наращивание новых случаев и примерно к

ноябрю достигается пик. Однако, если количество заболеваний после этого идет только на спад, то количество вакцинаций продолжает расти и достигает новых пиков включая до декабря и лишь потом идет на спад.

Четвертая волна уже показывает, что лаг между количеством заболеваний и количеством вакцинированных сошел на нет, но все же разница в продолжительности пика есть и составляет примерно 1 месяц.

Закончим последней волной, которая была в начале 2022 года:





Важно заметить, что случаи вакцинированных 5 волны колеблются в районе 0-3000 человек(на миллион) в сутки, тогда как цифры в остальных были в районе 2000-6000 человек(на миллион) в сутки. Т.е. цифры вакцинированных 5 волны в разы меньше в то время как цифры заболевших больше. Вероятно, количество вакцинированных

уменьшается, потому что большая часть людей уже переболели или вакцинировались. Однако графики все равно очень схожи и пик достигается в конце феврале, а дальше идет на спад.

Изучив графики вакцинированных и заболевших, можно сделать вывод, что они примерно схожи и пик заболевших приходятся на пики вакцинированных, хотя и с небольшим опозданием.

3. Дисперсионный анализ

Первым серьезным методом, который я буду использовать станет дисперсионный анализ. Я решил изучить насколько разные данные по коронавирусу зависят от страны. Т.е. я решил сравнить данные по странам и найти государства, где вирус вел себя одинаково.

Для анализа я буду использовать однофакторный дисперсионный анализ и критерий Тьюки. Нулевой гипотезой для всех данных будет то, что данные по коронавирусу ведут себя одинаково вне зависимости от страны, т.е. в двух странах данные не различаются. Уровень значимости составит $\alpha=0.05$. Если удастся отвергнуть нулевую гипотезу, т.е. $\alpha<0.05$, то это значит, что дислокация вируса(в нашем случае страна) влияет на интересующую нас переменную.

Для начала я решил изучить данные по новым заболевшим среди стран Европы. Для этого я использовал критерий Тьюки и отобрал все пары стран, в которых p-значение больше 0.05, т.е. нулевая гипотеза выполняется. Получилось 358 пар[1] стран. Посмотрим на пары стран, одной из которых является Россия. Так, Болгария, Финляндия, Венгрия, Мальта, Молдова, Польша, Румыния, Швеция, Ватикан имеют схожее поведение с Россией.

Дальше я посмотрел на страны[2], которые имеют схожие темпы новых вакцинаций с Россией. Этими странами оказались Беларусь, Хорватия, Эстония, Латвия, Молдова, Северная Македония, Польша, Румыния, Словакия, Словения, Сербия. Заметим, что в списках заболевших и вакцинированных есть лишь две страны: Молдова и Румыния. Также среди этих стран лишь Беларусь, Молдова, Северная Македония, Сербия и Словакия одобрили вакцину Спутник-V.

Далее я взял топ-10 стран по ВВП(ППС) и посмотрел на данные[3] по новым случаям заболеваний. Схожие с Россией данные имели такие страны как США, Япония, Индонезия, Индия, Китай и Бразилия. Лишь страны Европы не вошли в этот список.

[1] Все 358 пар можно посмотреть в файле ‘Europe_cases.html’ , приложенном к письму.
[2] Файл ‘Europe_vaccinations’.
[3] Файл ‘top_10_cases.html’.

Потом мне стало интересно, какие страны Европы имеют схожесть со странами не из Европы. Оказалось, что есть лишь 3 пары: Англия-Бразилия, США-Германия, США-Англия.

Если же рассматривать эти же страны только на данных[4] по новым вакцинациям, то лишь 8 пар, где нулевая гипотеза нарушается и из этих 8 пар в шести присутствует Россия, а в остальных двух - Япония. Т.е. почти во всех странах попарно(кроме России) примерно одинаковые данные по новым вакцинациям, что говорит о том, что в топ-10 странах по ВВП(ППС) количество новых вакцинаций вело себя схожим образом.

4. Корреляция и линейная регрессия

Последним шагом станет изучение данных с помощью корреляции и линейной регрессии. Забегая вперед, очевидных взаимосвязей и интересных фактов обнаружить я не смог, но решил вставить эту часть, чтобы показать, что каких-то связей просто нет.

Снова проанализируем топ-10 стран по ВВП на наличие корреляций по новым заболеваниям, новым вакцинациям и смертям. Заодно проверим, насколько верно мы сделали вывод при дисперсионном анализе.

Первое, получим коэффициент корреляции(Пирсона) по новым случаям заболеваний. Напомню, что при дисперсионном анализе России мы получили, что лишь страны Европы отличались от России. Самые маленькие коэффициенты получили у США, Китая и Индии – 0.11, -0.11 и -0.026 соответственно. Самый большой у Индонезии – 0.60, что говорит о средней корреляции. У остальных стран он находится в промежутке между 0.27 и 0.45, что означает о слабой корреляции. Также важно заметить, что отрицательная корреляция лишь с Китаем и Индией – странами из Азии.

Теперь посмотрим на данные по новым вакцинациям. Самые маленькие коэффициенты получили у США и Японии – 0.043 и 0.075 соответственно. Все остальные страны имели коэффициент в промежутке между 0.22 и 0.42, что говорит о слабой корреляции. Стран с сильной корреляцией не выявлено. Одной из причин этому может быть то, что Спутник-V был основной вакциной лишь в РФ.

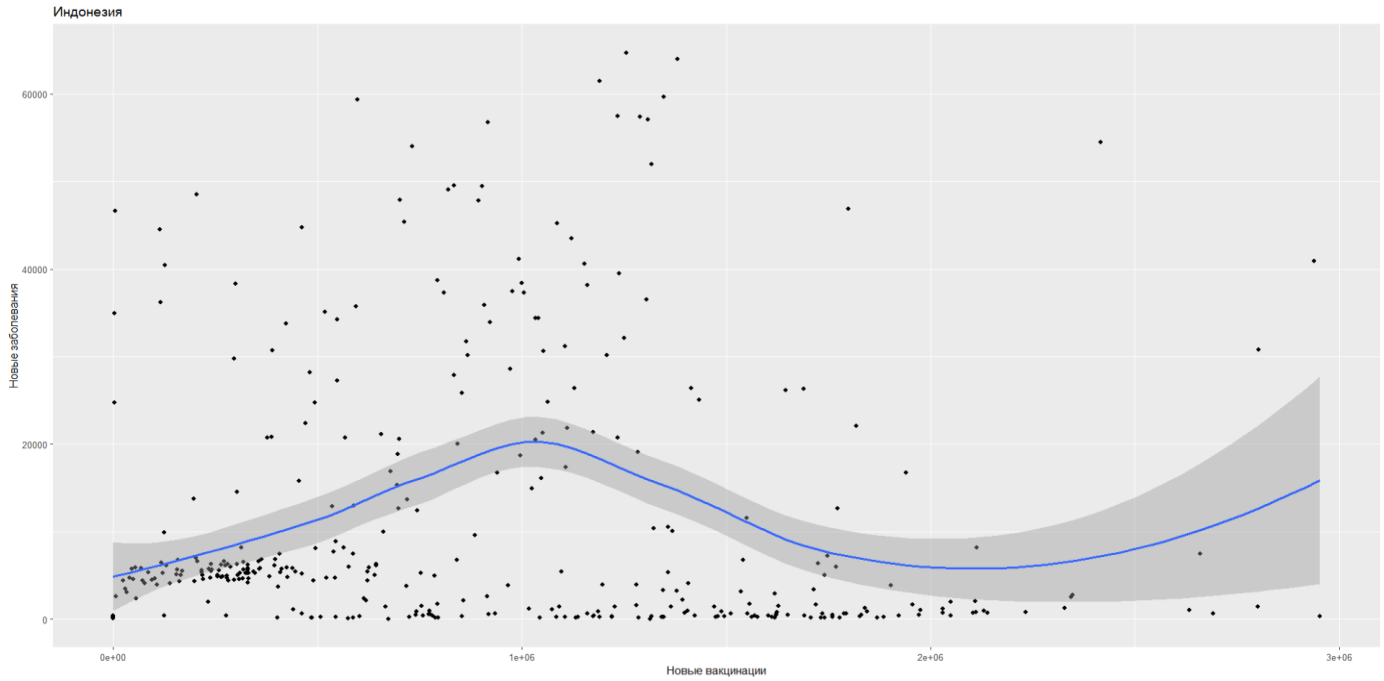
[4] Файл ‘top_10_vaccinations.html’.

Со смертями дело обстоит несколько иначе. Почти все страны, кроме США имеют очень низкую корреляцию. У США же коэффициент равен 0.43, что говорит о слабой-средней корреляции.

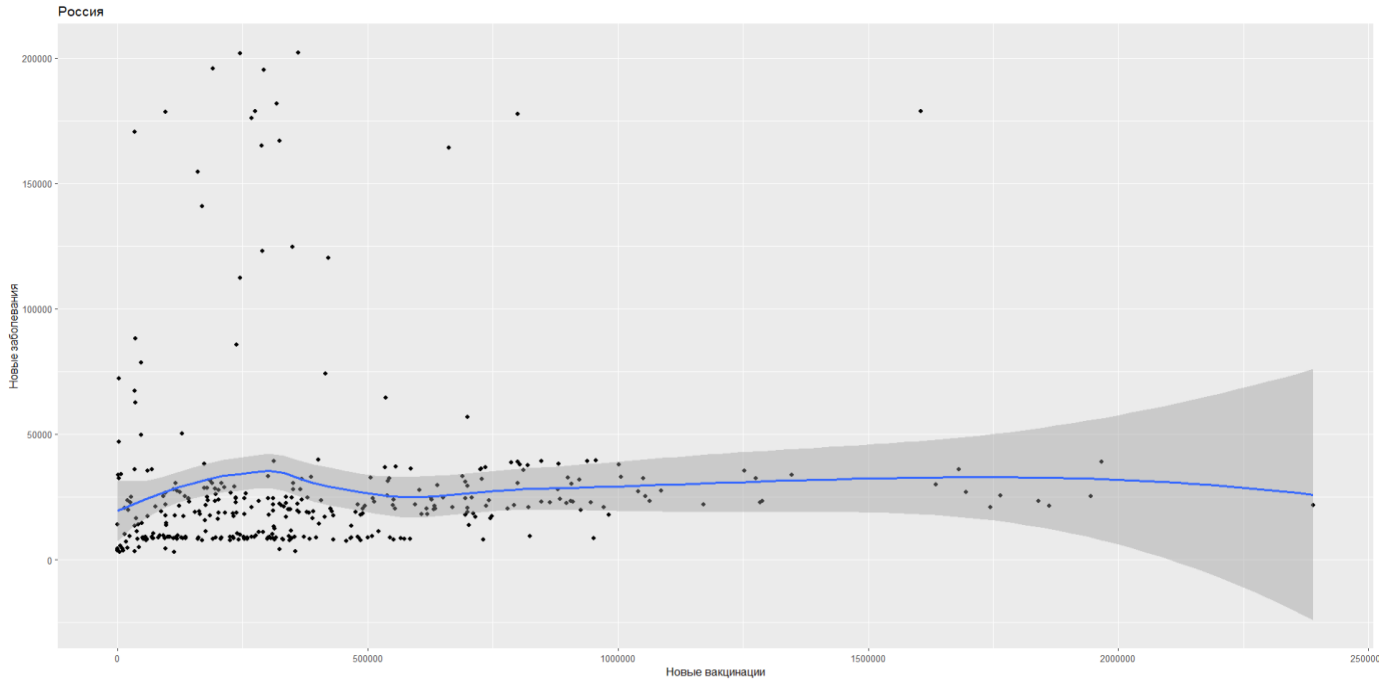
Теперь перейдем к линейно регрессии, однако будем смотреть на взаимосвязь новых вакцинаций и заболеваний в каждой стране отдельно.

Я начертил диаграмму рассеяния для каждой страны и обнаружил, что графики очень схожи между собой. Покажу лишь два из них, остальные ровно такие же.

Индонезия



Россия



Как видно, они имеют очень похожие линии тренда, точки экстремума и в целом ведут себя схоже.

Из диаграммы рассеяния видно, что данные имеют очень много выбросов и в целом не наблюдается линейной зависимости, что говорит о том, что линейную регрессию строить будет не разумно. Чтобы проверить это, посчитаем R^2- долю объясненной дисперсии- и убедимся, что он близок к нулю, т.е. верно будет сделать вывод об отсутствии линейной взаимосвязи.

Посчитаем данные показатели с помощью функции **lm()**. Сделав это для всех десяти и стран, получаем среднее значение на уровне 0.05, что и позволяет нам сделать вывод об отсутствии линейной взаимосвязи.

5. Заключение

Каких-то гениальных выводов или открытий из этой работы сделать нельзя, все-таки знаний в математической статистике и в целом в R не так много, но я хотел показать, что и с помощью обычных методов можно попытаться изучить данные, сделать какие-то выводы и получить знания об изучаемом объекте. Несмотря на некоторую сухость и простоту в некоторых случаях, я все же хочу подчеркнуть, что я только еще начинаю свой путь в анализе данных и мне предстоит многое узнать и многому научиться, в чем я и надеюсь, мне поможет Центр математических финансов.