**Документация**

**Дипломная работа по теме:**

**“Анализ распространения COVID-19 (поиск инсайтов, составление рекомендаций стейкхолдерам, построение предиктивной модели)”**

**Профессия “Аналитик данных”, DAU-27**

**Гарипов Алмаз Русланович**

**г. Москва, 2023**

**Оглавление**

**Введение…………………………………………………………………………….3**

**Блок 1. Описание исходного датасета и типов данных……4**

**1.1 Датасет для проекта ……………………………………………………..4**

**1.2 Таблица confirmed\_global……………………………………………..4**

**1.3 Таблица deaths\_global…...………………………………………………4**

**Блок 2. Подготовка и преобразование данных….……………..5**

**2.1 Подготовка..……………….…………………………………………………5**

**2.2 Добавление дополнительных атрибутов..……………….…….5**

**2.3 Преобразование..……….…….…..……………………………………….6**

**Блок 3. Очистка Данных……..……………………………………………...6**

**3.1 Аномалии в данных………………………….…………………………...7**

**Блок 4. Анализ данных для стейкхолдеров……………………….7**

**4.1 Анализ общих показателей…………………………………………….7**

**4.1.1 Динамика зараженных COVID-19……………….…………..………….7**

**4.1.2 Динамика смертности от COVID-19………………………………...7**

**4.1.3 Динамика по процентному соотношению погибших к**

**заболевшим COVID-19………………………………………….…………..…8**

**4.1.4 Динамика заболеваемости в разрезе континентов……………9**

**4.2 Топы по показателям заболеваемости в мире на текущий**

**момент.…………………………………………………………………..………..10**

**4.2.1 Топ- 20 стран по количеству случаев заражения по**

**состоянию на март 2023г.………………………………………………..…10**

**4.2.2 Топ-20 стран по количеству погибших по состоянию на**

**март 2023г…………………………………………………………………………………….11**

**Блок 5. Построение модели Линейной Регрессии…………...12**

**5.1 Оценка качества прогнозной модели (R^2)……………..12**

**5.2 Оценка качества прогнозной модели(RMSE).....................12**

**5.3 Визуализация итогов предсказания модели……….………..12**

**Итоги проекта и заключение……………………………………….......13**

**Введение**

**Цели проекта:**

В рамках работы в Крупной фармацевтической компании, занимающейся реализацией и продвижением вакцины от COVID-19 по всему Миру, провести анализ имеющихся данных с целью определения самых востребованных регионов и трендов развития заболеваемости в Мире для составления рекомендаций по продвижению товара и стратегии развития компании

**Бизнес-Задачи:**

1. Определение наиболее привлекательных регионов для развертывания производств в целях сокращения логистических затрат;

2. Выявление регионов – основных заказчиков продукции;

3. Анализ основных Тенденций развития заболеваемости в целях определения жизненного цикла продукции для последующего бюджетирования проекта.

**Блок 1. Описание исходного датасета и типов данных**

**1.1 Датасет для проекта**

Для исследования был взят датасет “[**COVID-19 Time Series Data**](https://data.world/shad/covid-19-time-series-data)” с данными по Covid-19

**1.2 Таблица confirmed\_global**

|  |  |  |  |
| --- | --- | --- | --- |
| **№** | **Имя Столбца** | **Описание** | **Тип данных** |
| ***1*** | ***Province/State*** | Провинция/штат | object |
| ***2*** | ***Country/Region*** | Страна/регион | object |
| ***3*** | ***Lat*** | Широта | float64 |
| ***4*** | ***Long*** | Долгота | float64 |
| ***5*** | ***1/22/20*** | Кол-во подтвержденных случаев заболеваний на дату 22.01.2020 | int64 |
| ***6*** | ***1/23/20*** | Кол-во подтвержденных случаев заболеваний на дату 23.01.2020 | int64 |
| ***7*** | ***1/24/20*** | Кол-во подтвержденных случаев заболеваний на дату 24.01.2020 | int64 |
| ***…*** | ***…*** | … | … |
| ***1147*** | ***3/9/23*** | Кол-во подтвержденных случаев заболеваний на дату 09.03.2023 | int64 |

**1.3 Таблица deaths\_global**

|  |  |  |  |
| --- | --- | --- | --- |
| **№** | **Имя Столбца** | **Описание** | **Тип данных** |
| ***1*** | ***Province/State*** | Провинция/штат | object |
| ***2*** | ***Country/Region*** | Страна/регион | object |
| ***3*** | ***Lat*** | Широта | float64 |
| ***4*** | ***Long*** | Долгота | float64 |
| ***5*** | ***1/22/20*** | Кол-во умерших на дату 22.01.2020 | int64 |
| ***6*** | ***1/23/20*** | Кол-во умерших на дату 23.01.2020 | int64 |
| ***7*** | ***1/24/20*** | Кол-во умерших на дату 24.01.2020 | int64 |
| ***…*** | ***…*** | … | … |
| ***1147*** | ***3/9/23*** | Кол-во умерших на дату 09.03.2023 | int64 |

**Блок 2. Подготовка и преобразование данных**

В ходе исследования качества данных были сделаны следующие

изменения:

**2.1 Подготовка**

* Обе таблицы были транспонированы с помощью метода melt - столбцы с датами были перенесены в строки, а показатели с количеством подтвержденных случаев заболеваний/смертей вынесены в столбцы. Данная операция была выполнена с целью приведения данных в нормальную форму, т.к. дальнейшая работа с датафреймом с 1147 столбцами была бы затруднительна.
* После нормализации данных два датафрейма были объединены с помощью метода merge по следующим столбцам: Province/State','Country/Region','date', 'Lat', 'Long’, использовался способ слияния inner (проверка на количество строк показала, что мы не потеряли данные при «внутреннем» слиянии). Получаем объединенный датафрейм:

|  |  |  |  |
| --- | --- | --- | --- |
| **№** | **Имя Столбца** | **Описание** | **Тип данных** |
| ***1*** | ***Province/State*** | Провинция/штат | object |
| ***2*** | ***Country/Region*** | Страна/регион | object |
| ***3*** | ***Lat*** | Широта | float64 |
| ***4*** | ***Long*** | Долгота | float64 |
| ***5*** | ***Date*** | Дата | object |
| ***6*** | ***Confirmed*** | Кол-во подтвержденных случаев заболеваний | int64 |
| ***7*** | ***Deaths*** | Кол-во умерших | int64 |

**2.2 Добавление дополнительных атрибутов**

* В целях добавления вариативности работы с данными и возможности дополнительных верхнеуровневых визуализаций был добавлен атрибут, который содержит в себе континент страны.

**2.3 Преобразование**

* В столбце “***Date***” - Изменен тип данных на “ **datetime64[ns]**”

**Блок 3. Очистка данных**

Результат итогового вида обработанного датасета, а также выводы по причинам очистки данных приведены в таблице:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **№** | **Имя Столбца** | **Преобразование**  **данных** | **% NaN** | **Очистка**  **данных** |
| ***1*** | ***Date*** | datetime64[ns] | 0.00 | Без изменений, нет пропущенных значений |
| ***2*** | ***Continent*** | object | 0.00 | Были удалены все объекты «Not Continent». Исходный датасет содержит в себе данные не только по странам, но и по завершившимся событиям (Олимпийские игры 2020/2022, лайнеры Diamond Princess и MS Zaandam) – в рамках моделирования такие данные не потребуются. |
| ***3*** | ***Country/Region*** | object | 0.00 | Без изменений, нет пропущенных значений |
| ***4*** | ***Province/State*** | object | 68.51 | Пропущенные значения заменены на название страны/региона |
| ***5*** | ***Lat*** | float64 | 0.69 | Пропущенные значения заменены на координаты, которые чаще всего встречаются в рамках страны |
| ***6*** | ***Long*** | float64 | 0.69 | Пропущенные значения заменены на координаты, которые чаще всего встречаются в рамках страны |
| ***7*** | ***Confirmed*** | int64 | 0.00 | Без изменений, нет пропущенных значений |
| ***8*** | ***Deaths*** | int64 | 0.00 | Без изменений, нет пропущенных значений |

**3.1 Аномалии в данных**

Аномалии в данных не обнаружены.

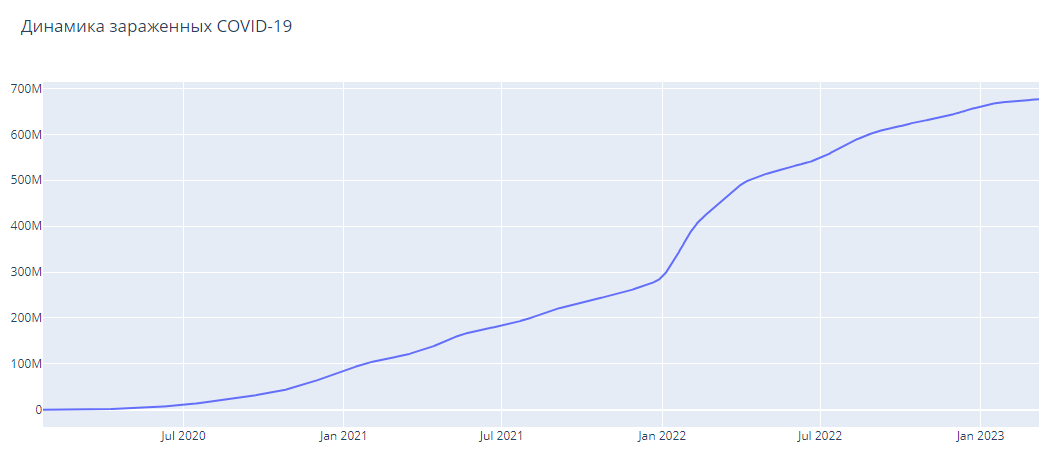
**Блок 4. Анализ данных для стейкхолдеров**

Целью блока является поиск тенденций и инсайтов для

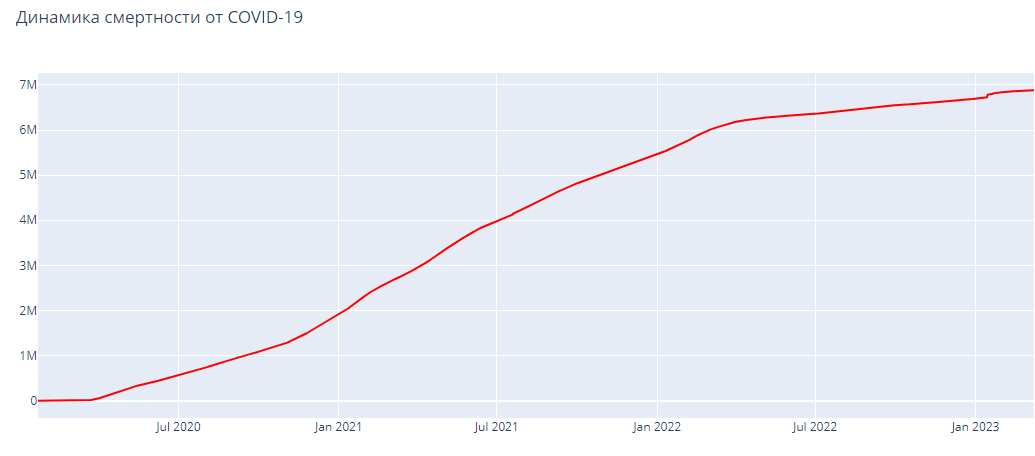
составления рекомендаций стейкхолдерам.

**4.1 Анализ общих показателей**

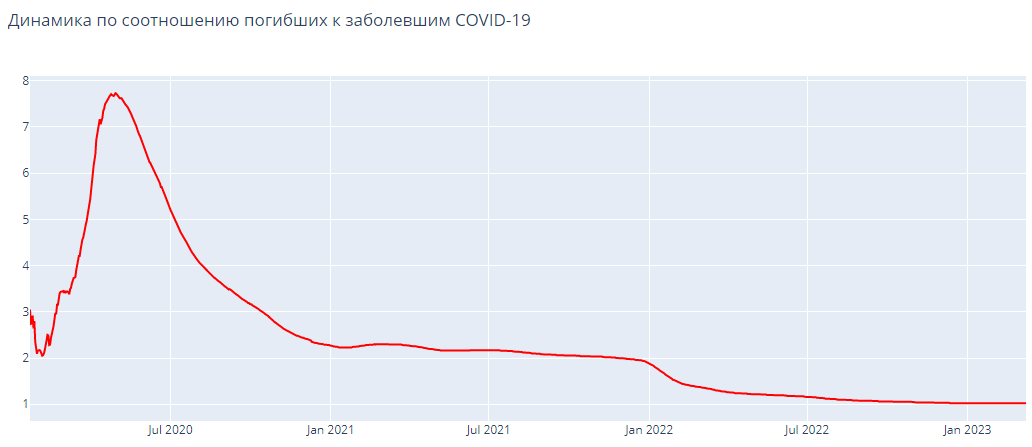
**4.1.1 Динамика зараженных COVID-19**



**4.1.2 Динамика смертности от COVID-19**



**4.1.3 Динамика по процентному соотношению погибших к заболевшим COVID-19**



**Вывод:**

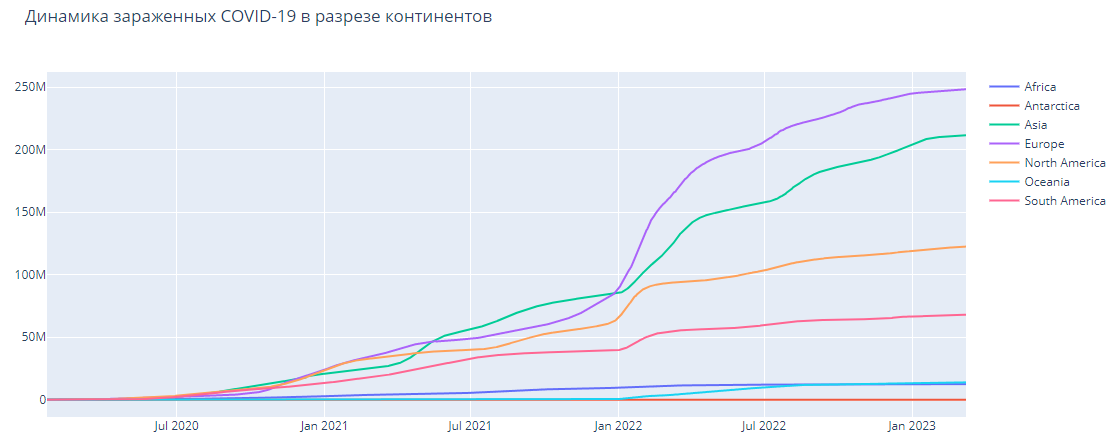
Появление COVID-19 значительно повлияло на мировую финансовую ситуацию и на модели развития компаний. Пандемия продолжается до сих пор, наблюдается увеличение количества подтвержденных случаев заболевания и количества смертей, однако, наблюдая на соотношение погибших к заболевшим, можно смело утверждать, что мероприятия по сдерживанию пандемии работают (ограничение передвижения, вакцинация, перевод на удаленную работу).

На текущий момент количество летальных исходов составляет 1,02% от количества подтвержденных случаев заболевания.

**Рекомендация Стейкхолдерам:**

Владельцам бизнеса и руководителям необходимо учитывать темпы развития коронавирусной инфекции COVID-19 в своей стратегии по развертыванию и ведения бизнеса. Также при возможности продолжать переводить собственных сотрудников (в особенности ключевых) на Remote-работу.

**4.1.4 Динамика заболеваемости COVID-19 в разрезе континентов**



**Вывод:**

По количеству зараженных COVID-19 за весь период антилидерами топа являются Европа и Азия, а наименьшее количество подтвержденных случаев заболевания наблюдается в Антарктике, Африке и Океании

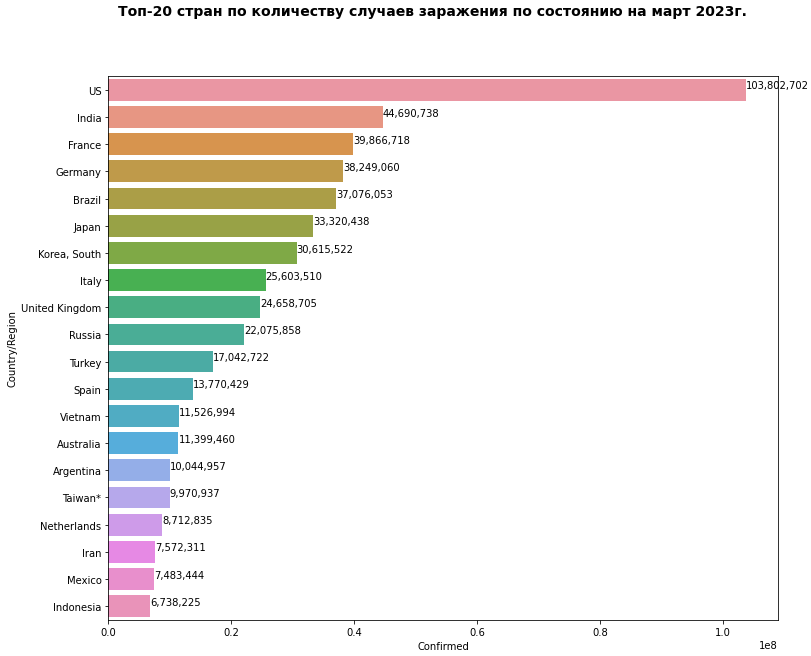
**Рекомендация Стейкхолдерам:**

Владельцам бизнеса, руководителям и менеджерам экспортного департамента, необходимо учитывать географическую составляющую развития COVID-19 и сфокусироваться на внедрении в рынки Европы, Азии и Северной Америки.

Для развертывания производств привлекательными локациями являются страны Южной Америки (последующий сбыт в Северную Америку), страны Африки (последующий сбыт в Европу и Азию).

**4.2 Топы по показателям заболеваемости в мире на текущий момент.**

**4.2.1 Топ- 20 стран по количеству случаев заражения по состоянию на март 2023г.**



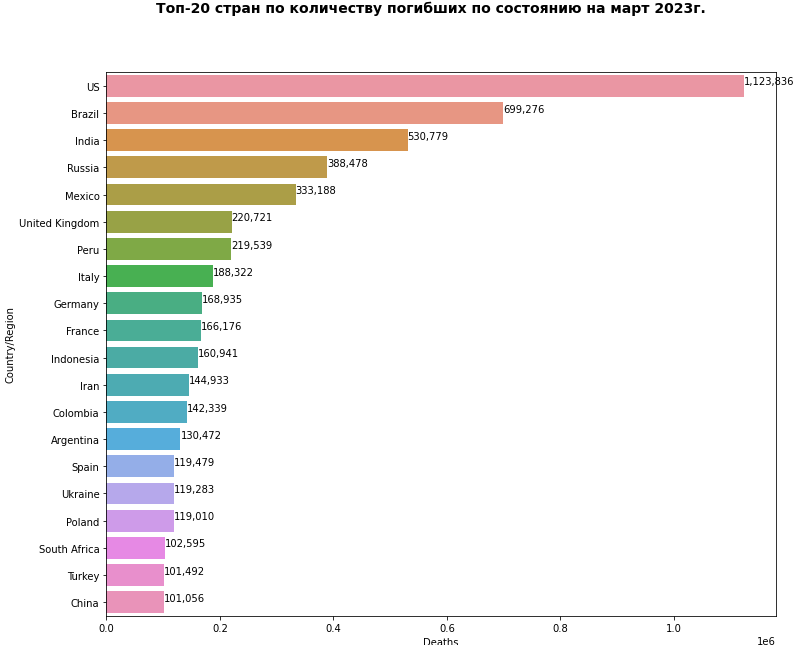
**Вывод:**

Получившийся топ стран отражает абсолютных «лидеров» по заболеваемости COVID-19 в мире.

**Рекомендация Стейкхолдерам:**

* Топ можно использовать для определения основных потенциальных покупателей продукции компании;
* Менеджерам из маркетингового департамента рекомендуется увеличить активность продвижения продукции в США, т.к. по количеству заболеваний, они являются абсолютными «лидерами»

**4.2.2 Топ-20 стран по количеству погибших по состоянию на март 2023г.**



**Вывод:**

Помимо стран-«лидеров» по количеству заболеваний в абсолютном выражении, также стоит обратить внимание на страны с высоким показателем смертности, например, Бразилия.

Т.к. одной из целей существования компании, является не только генерирование финансового результата, но и социальная ответственность перед общественностью

**Блок 5. Построение модели Линейной Регрессии**

Целью блока является построение модели линейной Регрессии для

прогноза заболеваемости в течение года.

**5.1 Оценка качества прогнозной модели (R^2)**

В ходе оценки модели, был получен коэффициент детерминации равный 0.95 (как в выборке для обучения, так и в выборке для теста). Для приемлемых моделей предполагается, что коэффициент детерминации должен быть хотя бы не меньше 0.5 (в этом случае коэффициент множественной корреляции превышает по модулю 0.7). Модели с коэффициентом детерминации выше 0.8 можно признать достаточно хорошими (коэффициент корреляции превышает 0.9%).

**5.2 Оценка качества прогнозной модели (RMSE)**

В ходе оценки были получены результаты, что в среднем наша

модель ошибается на 57.7 миллионов единиц с выборкой для

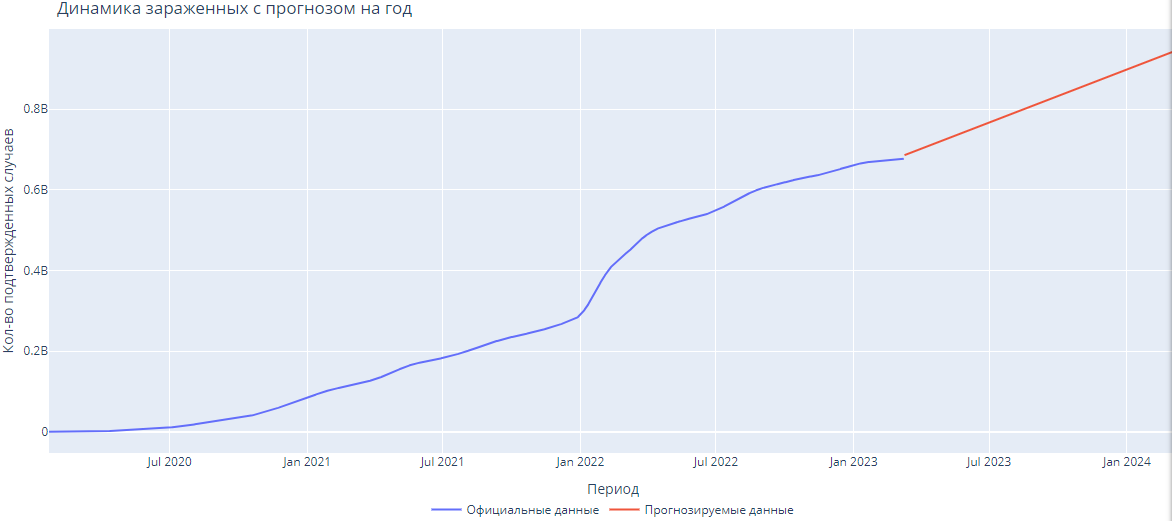
обучения и на 58.3 миллионов единиц с выборкой для теста,

поэтому качество и погрешность приемлемы, учитывая малый

объем данных, который мы подаем модели. Это говорит о

работоспособности модели для данного исследования.

**5.3 Визуализация итогов предсказания модели**



**Вывод:**

Модель показывает нам, что ожидается дальнейшее развитие заболеваемости COVID-19 и к 8 марта 2024 года ожидается 945 млн. подтвержденных случаев заболевания.

**Итоги проекта и заключение**

Сегодня мировая ситуация с COVID-19 находится на стадии продолжения развития заболеваемости и для владельцев фармацевтического бизнеса были найдены нужные инсайты и рекомендации, которые смогут помочь им в построении стратегии на будущие годы и улучшения своих показателей.

**По бизнес-задачам:**

1. Для задачи определения наиболее привлекательных регионов для развертывания производств в целях сокращения логистических затрат были определены относительно безопасные регионы базирования производств.

2. Были выявлены регионы – основные заказчики продукции, они же ТОП-страны по количеству заболеваний с трендом на дальнейшее увеличение заболеваемости.

3. Был проведен анализ основных тенденций развития заболеваемости, с помощью модели линейной регрессии, который показал, что спрос на продукт будет увеличиваться, как минимум год. Поэтому компания может вкладываться в дальнейшее расширение, увеличивать бюджет, как на производство, так и на маркетинг.