

Анализ данных о глобальном распространении Covid-19

Екимова Яна Сергеевна

Приамурский государственный университет имени Шолом-Алейхема

Студент

Аннотация

Данная научная статья посвящена проведению анализа данных о глобальном распространении Covid-19 с использованием Google Colaboratory. Для достижения поставленной цели использовался набор данных, содержащий подробную информацию о пандемии COVID-19. Используя Python и библиотеки анализа данных, такие как Pandas, Plotly.express был проведен анализ данных.

Ключевые слова: Google Colaboratory, Pandas, анализ данных.

Analysis of data on the global spread of Covid-19

Ekimova Yana Sergeevna

Sholom-Aleichem Priamurskiy State University

Student

Abstract

This scientific article is devoted to the analysis of data on the global spread of Covid-19 using Google Colaboratory. To achieve this goal, a dataset containing detailed information about the COVID-19 pandemic was used. Using Python and data analysis libraries such as Pandas, Plotly.express, data analysis was performed.

Key words: Google Coollaboratory, Pandas, data analysis.

1. Введение

1.1. Актуальность

Google Colab остается актуальным и популярным выбором для проведения исследований, разработки моделей машинного обучения и выполнения вычислений в облаке. В настоящее время COVID-19 является одной из наиболее серьезных и актуальных проблем в области здравоохранения, с которой сталкивается весь мир. Анализ данных о глобальной тенденции распространения вируса является неотъемлемой частью борьбы с пандемией и помогает лучше понять и прогнозировать динамику заболеваемости и принять соответствующие меры для сдерживания вируса.

1.2 Обзор исследований

А.Р.Сухинина, Е.В.Королева, Е.А.Ефимов, Н.А. Благой в исследовании продемонстрировали зависимость доступности жилья от объемов ввода

жилья. Темпов прироста населения, суммарного коэффициента рождаемости, коэффициента брачности, уровня занятости, удельного веса числа семей, получивших жилые помещения и улучшивших жилищные условия, количества предоставленных ипотечных жилищных кредитов, средневзвешенного срока и средневзвешенной процентной ставки по ипотечным жилищным кредитам [1]. Н.Е. Косых предложил подход к подбору наиболее производительного набора параметров для объекта классификатора текста. Для вычислений использовал облачный сервис Google Colaboratory, выполняющий код на языке Python внутри браузера, используя виртуальные аппаратные ресурсы [2]. Т.М. Татарникова, Е.Д.Пойманова, П.Ю.Богданов, Е.В.Краева, С.А. Веревкин в статье рассмотрели способы и методы изучения и построения нейронных сетей. Показано, что изучение принципов функционирования нейронных сетей, их применение для решения тех или иных задач возможны только через практику. Проведен анализ различных программных сред, которые могут быть использованы на лабораторных и практических занятиях по изучению и применению нейронных сетей. Выделен современный облачный сервис Google Colaboratory, рекомендуемый для обучения основам нейронных сетей благодаря наличию в нем предустановки библиотеки Tensorflow и библиотеки для работы на языке Python, бесплатного доступа к графическим процессорам, возможности написания и выполнения программного кода в браузере, а также отсутствию необходимости специальной настройки сервиса [3]. Е.А.Григорьев, Н.С.Климов в статье рассматривали разведочный анализ данных. Описали инструменты реализации анализа, библиотеки Python. Представлен пример выполненный на данных обнаружению присутствия людей в помещении [4].

1.3 Цель исследования

Цель исследования – выполнить анализ данных с помощью Google Colaboratory.

2. Материалы и методы

В данном исследовании используется Google Colaboratory для анализа данных.

Набор данных о глобальном распространении Covid-19 содержит информацию о распространении болезни по месяцам и годам в различных городах мира.

Данные для работы можно скачать по ссылке <https://www.kaggle.com/code/nelgiriyeewithana/an-introduction-to-global-covid-trend-dataset>.

3. Результаты

Создадим блокнот в Google Colab и загрузим файл "covid19.csv".

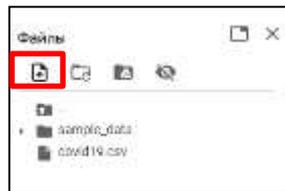


Рис.1-Загрузка данных

В блоке кода импортируем необходимые библиотеки и модули для анализа данных, в данном случае Pandas, Plotly.express. Загружаем данные из CSV-файла (Рис.2).

```
[1]: import plotly.express as px
import pandas as pd

[2]: df = pd.read_csv('covid19.csv')
```

Рис.2-Код

Далее посмотрим глобальные случаи заболевания с течением времени.

```
global_cases = df[df.numeric_only=True].reset_index(name='Total Cases')
fig = px.line(global_cases, x='index', y='Total Cases', title='Global Cases Over Time')
fig.show()
```

Рис.3-Код

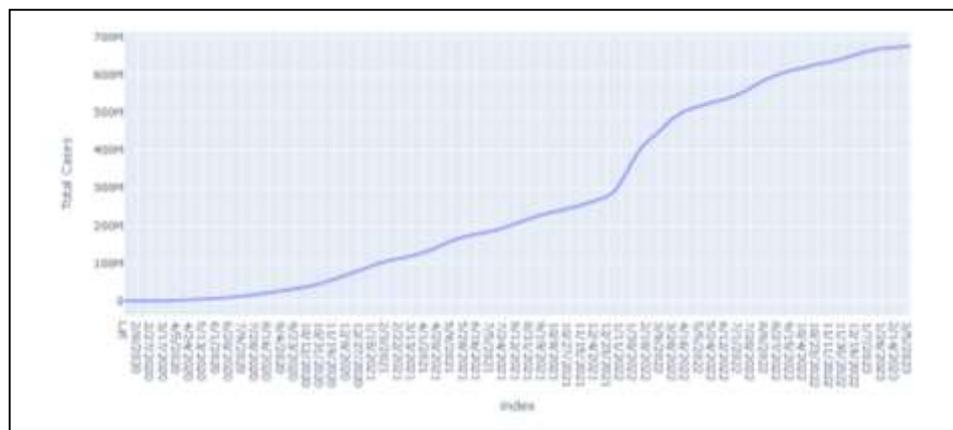


Рис.4-Итог

Теперь напишем код, чтобы посмотреть на ежедневные новые случаи заражения COVID-19.

```
[4]: daily_global_cases = df[df.numeric_only=True].diff().reset_index(name='Daily New Cases')
fig = px.line(daily_global_cases, x='index', y='Daily New Cases', title='Daily New Cases - Global')
fig.show()
```

Рис.5-Код

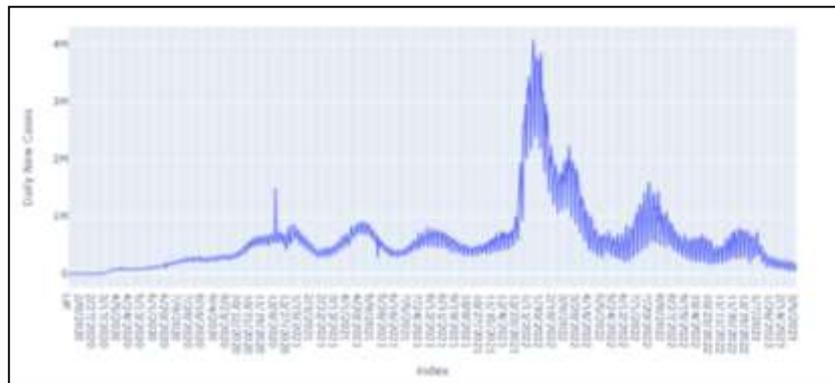


Рис.6-Ежедневные заражения

Далее посмотрим на ежедневные новые случаи заражения в разных континентах страны.

```
fig = px.line(df_melt(id_vars=['Country/Region'], var_name='date', value_name='Cases'),
             x='date', y='Cases', color='Country/Region',
             title='Cases Over Time - Selectable Countries')
fig.show()
```

Рис.7-Код

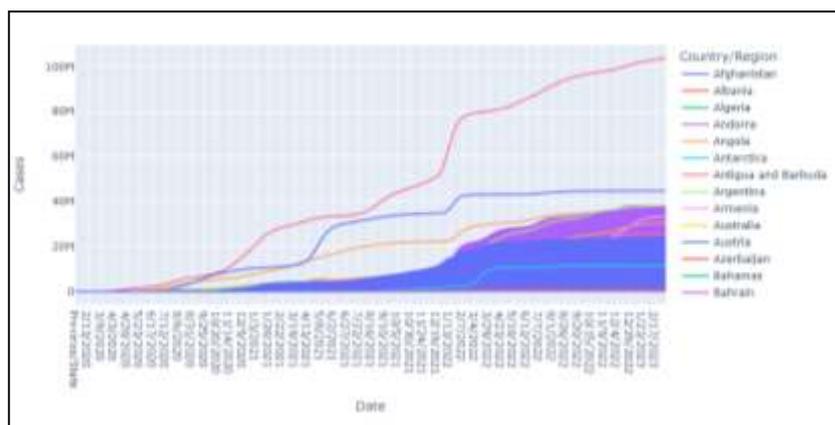


Рис.8-Случаи заражения по странам

Так же можно выбрать определенную страну или регион. В данном случае, на графике показаны только: Grenada, Guatemala, Guinea, Guyana, Haiti, India, Honduras, Iran.

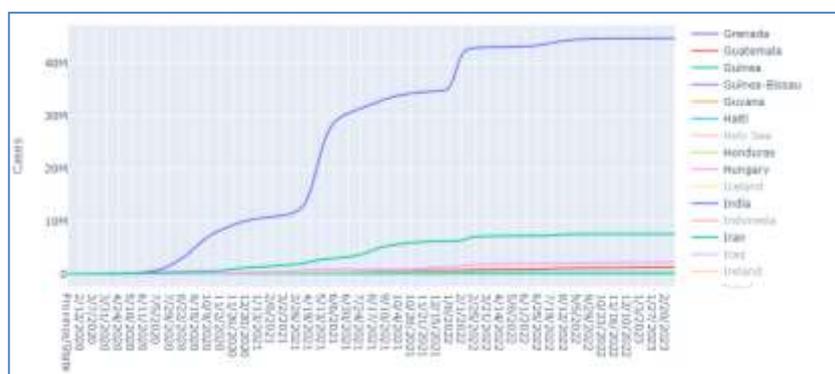


Рис.9-Случаи заражения по странам

Теперь напишем код, чтобы посмотреть общее количество случаев заражения по всему миру.

```
fig = px.choropleth(df, locations='Country/Region', locationmode='country names', color=df.columns[-1], title='Total Cases Worldwide')
fig.show()
```

Рис.10-Код

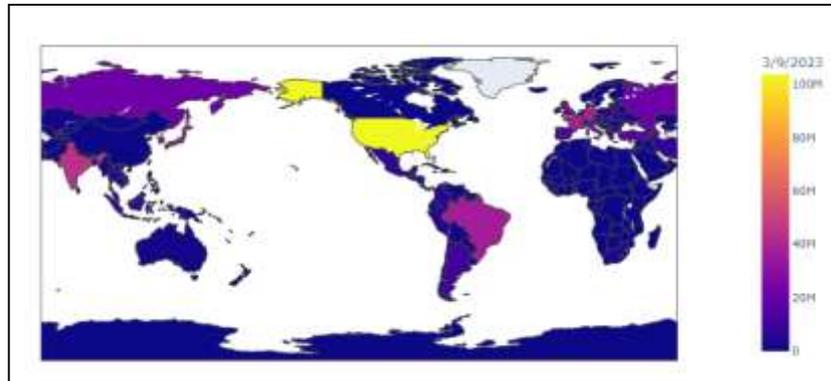


Рис.11-Общее количество заражения

Далее напишем код, чтобы посмотреть на случаи заражения в Азии (Рис.13) и по выбранным странам (Рис.15).

```
asia_cases_over_time = df[df['Country/Region'].isin(['China', 'India', 'Japan', 'South Korea'])].groupby('Country/Region').sum().T
fig = px.line(asia_cases_over_time, x=asia_cases_over_time.index, y=asia_cases_over_time.columns, title='Cases Over Time - Asia')
fig.show()

<ipython-input-7-0d000ed5a27>:1: FutureWarning:
The default value of numeric_only in DataFrameGroupBy.sum is deprecated. In a future version, numeric_only will default to False. To
```

Рис.12-Код

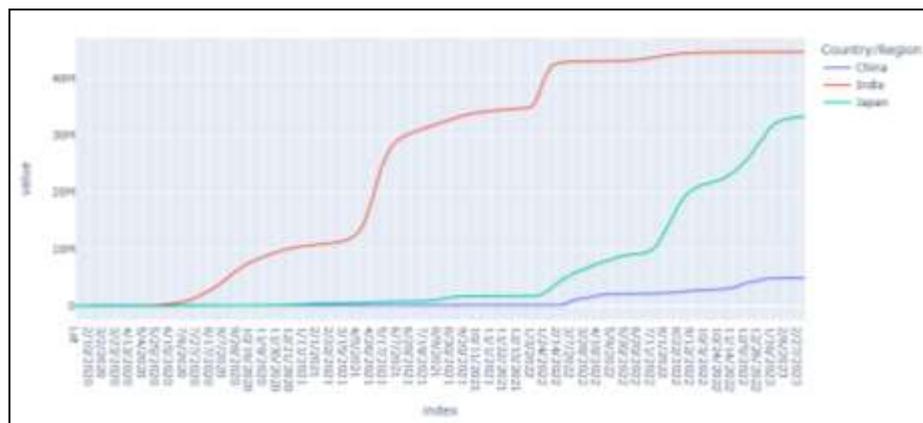


Рис.13-Случаи заражения в Азии

```
[8] selected_countries = ['United States', 'India', 'Brazil', 'Russia', 'United Kingdom']
selected_countries_cases_over_time = df[df['Country/Region'].isin(selected_countries)].reset_index().pivot(index='Date', columns='Country/Region', values='Cases')
fig = px.line(selected_countries_cases_over_time, x='Date', y='Cases', color='Country/Region', title='Cases Over Time - Selected Countries')
fig.show()
```

Рис.14-Код

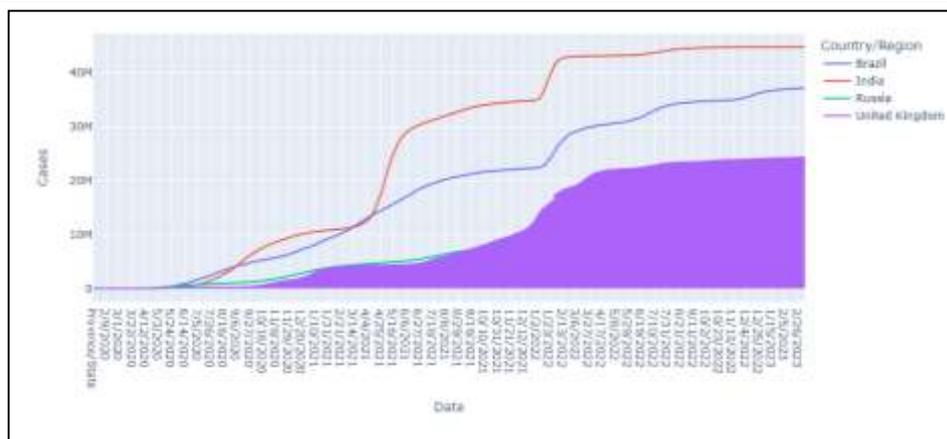


Рис.15- Количество заражения по выбранным странам

4. Выводы

В ходе исследования были выполнены следующие этапы: загрузка данных, визуализация данных с использованием графиков, анализ динамики распространения вируса в разных регионах.

Анализ данных о глобальной тенденции распространения COVID-19 в Google Colaboratory позволяет получить ценные инсайты для общества и принимающих решения в сфере общественного здравоохранения.

Библиографический список

1. Сухина А.Р., Королева Е.В., Ефимов Е.А., Благой Н.А. Доступность жилья и определяющие ее факторы в регионах России. 2022. С. 70-81.
2. Косых Н. Е. Оценка гиперпараметров при анализе тональности русскоязычного корпуса текстов//Интеллектуальные технологии на транспорте. 2020. №. 3 (23). С. 41-44.
3. Богданов П. Ю. и др. Программные среды для изучения основ нейронных сетей //Программные продукты и системы. 2021. №. 1. С. 145-150.
4. Григорьев Е. А., Климов Н. С. Разведочный анализ данных с помощью python //E-Scio. 2020. №. 2 (41). С. 165-176.