Projekt COVID

Link do danych: https://covid19datahub.io/

1. Wstęp

Celem projektu jest zapoznanie się z danymi związanymi z pandemią Covid-19 z platformy COVID-19 Data Hub (autorstwa Emanuele Guidotti i David Ardia) oraz sprawdzenie czy na podstawie zgormadzonych danych można dokonać globalnej i lokalnej predykcji liczby zachorowań oraz liczby śmierci spowodowanej chorobą COVID-19.

2. Podstawowa eksploracja danych:

Dane składają się z 47 kolumn i 286882 wierszy, które mają bardzo duże braki. Posługując się dokumentacją wybieramy wstępnie kolumny które mogą być użyteczne do analizy.
Dla wybranych zmiennych sprawdzamy średnią brakujących wartości:

date	0.00
confirmed	16.52
deaths	22.52
recovered	74.21
tests	69.38
vaccines	75.40
people_vaccinated	76.54
people_fully_vaccinated	77.37
hosp	83.83
icu	84.56
vent	97.60
school_closing	31.03
workplace_closing	31.03
cancel_events	31.04
gatherings_restrictions	31.04
transport_closing	31.04
stay_home_restrictions	31.04
internal_movement_restrictions	31.03
international_movement_restrictions	31.03
information_campaigns	31.03
testing_policy	31.03
contact_tracing	31.03
facial_coverings	30.95
vaccination_policy	30.92
elderly_people_protection	31.04
government_response_index	31.04
stringency_index	31.04
containment_health_index	31.04
economic_support_index	31.04
administrative_area_level_1	0.00
population	0.41
iso_alpha_3	0.14

W kolumnie administrative_area_level_1 są nazwy krajów jest ich 236.

Bez pogłębionej analizy usuwamy pozostałe kolumny z grupy Administrative areas, które mają 100% braków. Latitude i longitude nie są potrzebne do dalszej analizy. Z grupy danych ISO codes zostawiamy 1 kolumnę iso_alpha_3 możliwa korzyść przy wizualizacji (brak dla 3 krajów). Z grupy

danych External Keys usuwamy wszystkie klucze nie są one nam potrzebne do analizy mają duże braki i ciężko je przetworzyć.

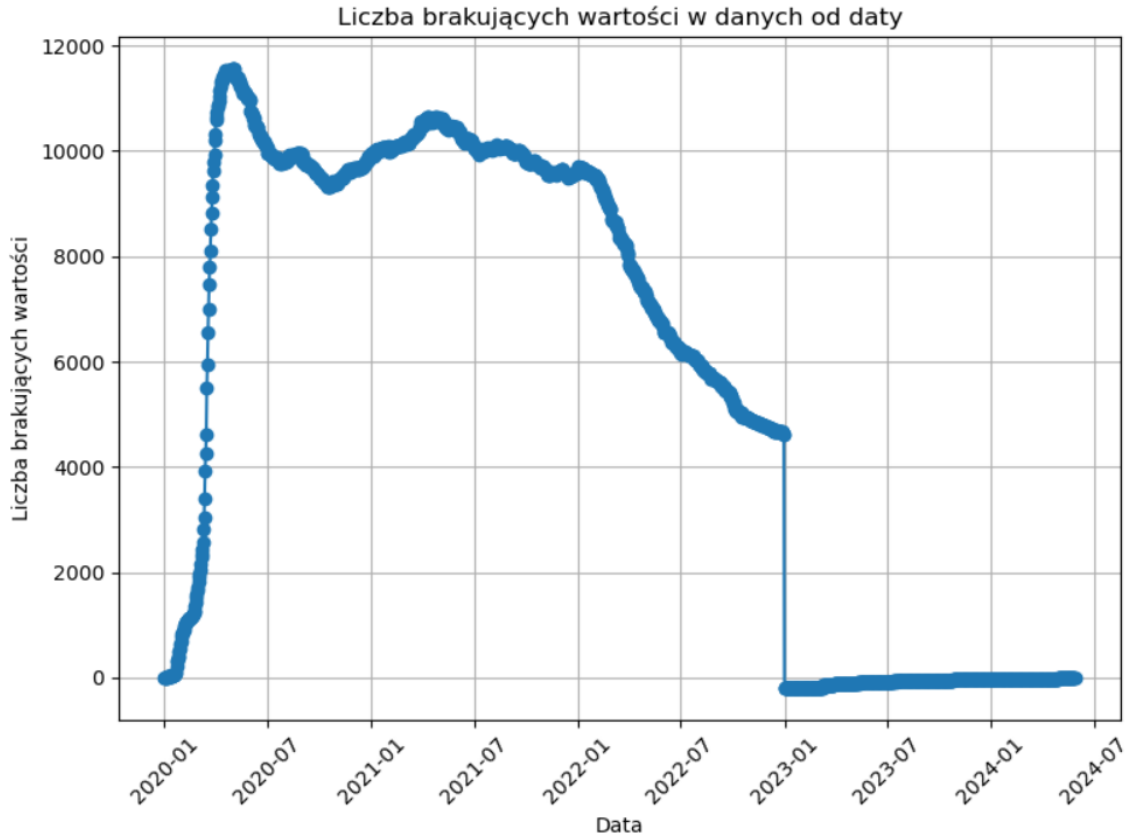
Zamieniamy format daty na bardziej użyteczny.

Po próbie odzyskania danych z kolumn 'recovered', 'tests', 'vaccines', 'people_vaccinated', 'people_fully_vaccinated', 'hosp', 'icu', 'vent' uznajemy je za nieistotne z powodu zbyt dużych braków danych i trudności w ich uzupełnieniu. Usuwamy dane z brakami ponad 65%.

Do dalszej analizy zostaje nam 24 kolumny.

Z powodu braku zakażeń po 5 maja 2023 usuwamy dalsze dane. Sprawdzam ilość danych dla konkretnych krajów. Usuwam kraje z ilością danych mniejsza niż półtorej ilości dni przez które były zbierane. Z powodu dużego braku danych w raportowaniu krajów od stycznie 2023 usuwamy dalsze dane.

```
Braki w kolumnie zakarzeń:
Oficjalny koniec pandemii:2023-05-05
Wszystkie dane: 16.52%
Przed 5.05.2023: 9.94%
Po 5.05.2023: 58.31%
```



Sprawdzamy ilość dni raportowanych przez każdy z krajów. Usuwamy kraje, które udostępniły jedynie dane z 50% danych względem najdłuższego okresu raportowania.

```
Diamond Princess 307
Colombia 655
Peru 677
Grand Princess 10
Costa Atlantica 81
Guernsey 695
Macao 692
Jersey 658
Pitcairn 441
Tokelau 533
```

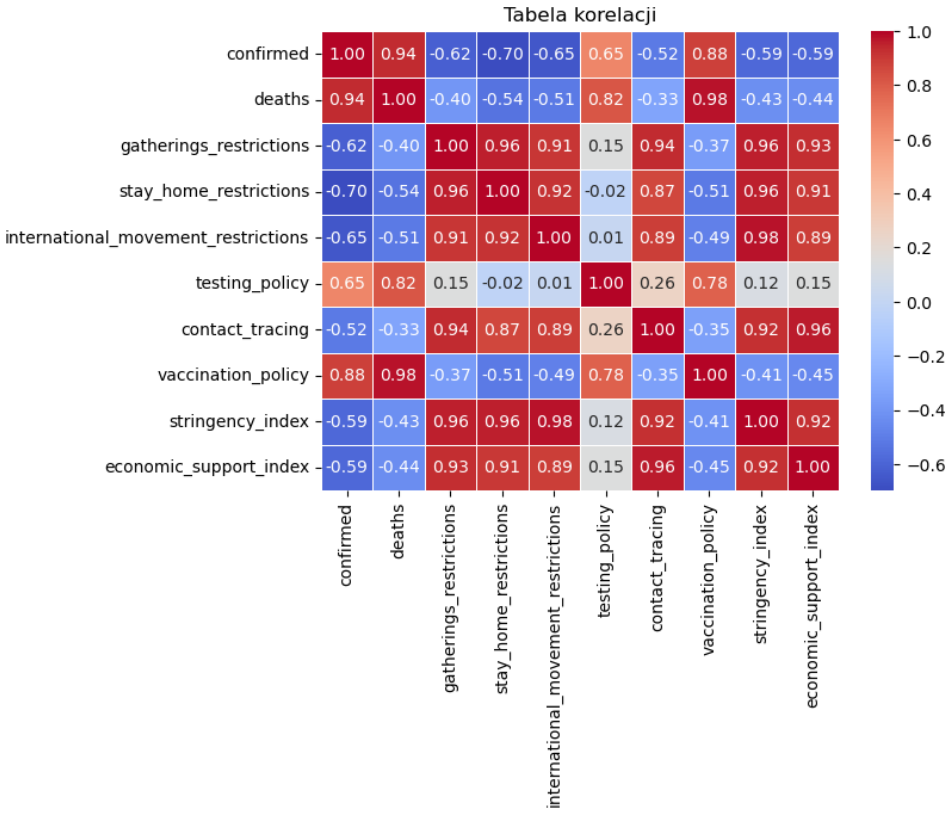
Sprawdzamy czy w danych są ujemne wartości. Zamieniamy ja na dodatnie.

3. Analiza globalna

W celu globalnej analizy sumujemy kumulatywne dane po dacie. Uzupełniamy przy tym braki w danych. Dla reszty kolumn podczas grupowania liczymy średnią wartości.

Usuwamy zmienne z najniższą korelacją: information_campaigns i facial_coverings. Z pierwszej grupy (zmienne o bardzo wysokiej korelacji między sobą) wybieramy trzy zmienne o najniższej korelacji między sobą i największej korelacji z zmiennymi objaśnianymi: gatherings_restrictions, stay_home_restrictions, international_movement_restrictions.Z drugiej grupy wybieramy dwie zmienne o najniższej korelacji między sobą i największej korelacji z zmiennymi objaśnianymi: stringency_index, economic_support_index.

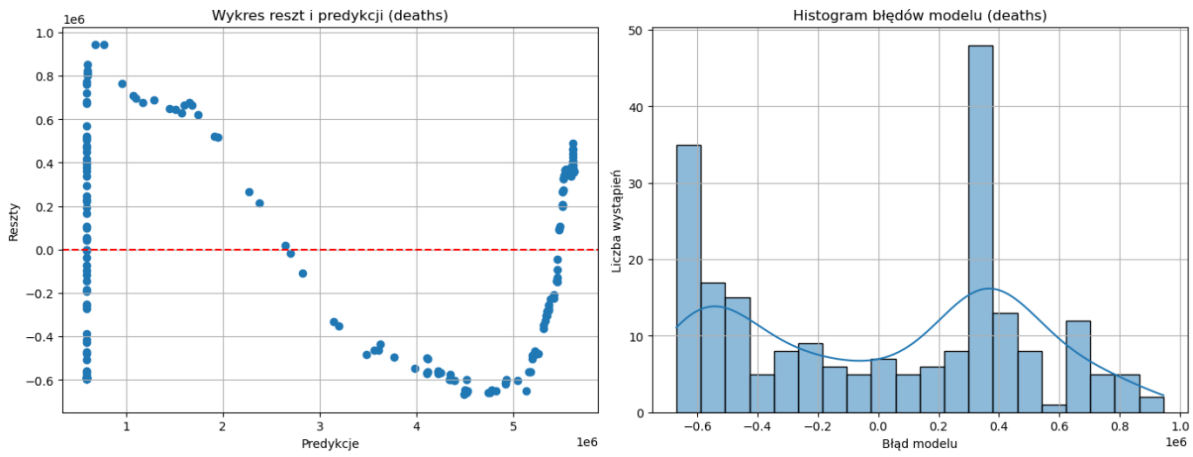
Dla wybranych zmiennych tworzymy tablice korelacji:

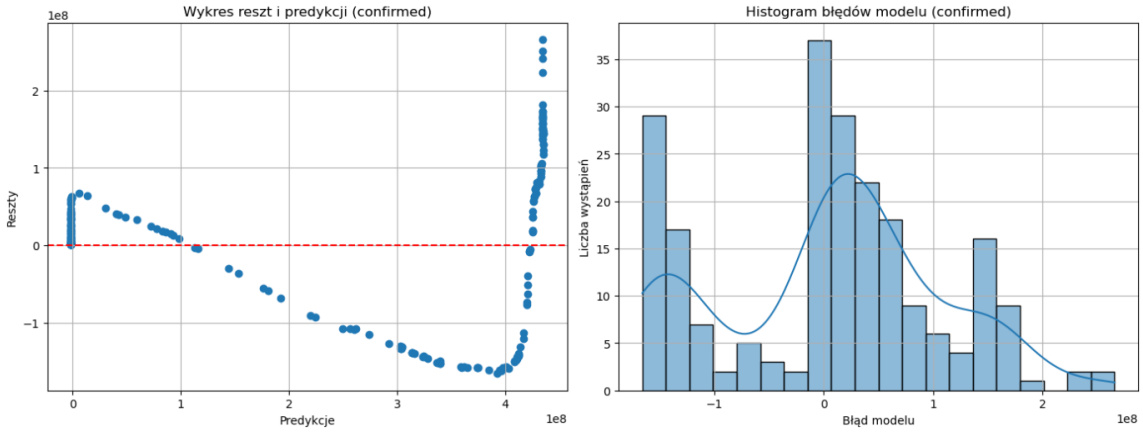


Wykonujemy regresję liniową od każdej zmiennej dla obu zmienny objaśnianych najlepsze metryki ma zmienna vaccination_policy

	objasniajaca	objasniana	mse	mae	r2
0	vaccination_policy	confirmed	10516096897410950.00	80512070.16	0.78
1	testing_policy	confirmed	27395638280713416.00	150392181.31	0.43
2	stay_home_restrictions	confirmed	27777837162650660.00	119099644.68	0.43
3	international_movement_restrictions	confirmed	28812911230459840.00	123176949.37	0.40
4	gatherings_restrictions	confirmed	32458086062834956.00	136037735.50	0.33
5	stringency_index	confirmed	34219653391678596.00	134496956.77	0.29
6	economic_support_index	confirmed	35502029493114172.00	127951653.45	0.27
7	contact_tracing	confirmed	36542699767533888.00	143707400.91	0.24
8	vaccination_policy	deaths	222236977436.95	429627.42	0.96
9	testing_policy	deaths	1667829543564.81	1144052.15	0.68
10	stay_home_restrictions	deaths	3990794334652.67	1599785.43	0.23
11	international_movement_restrictions	deaths	3993101659471.39	1604823.69	0.23
12	stringency_index	deaths	4494469923123.38	1743534.99	0.13
13	economic_support_index	deaths	4585875103316.17	1697808.34	0.11
14	gatherings_restrictions	deaths	4590636546310.13	1807423.72	0.11
15	contact_tracing	deaths	4729638859918.06	1841130.02	0.09

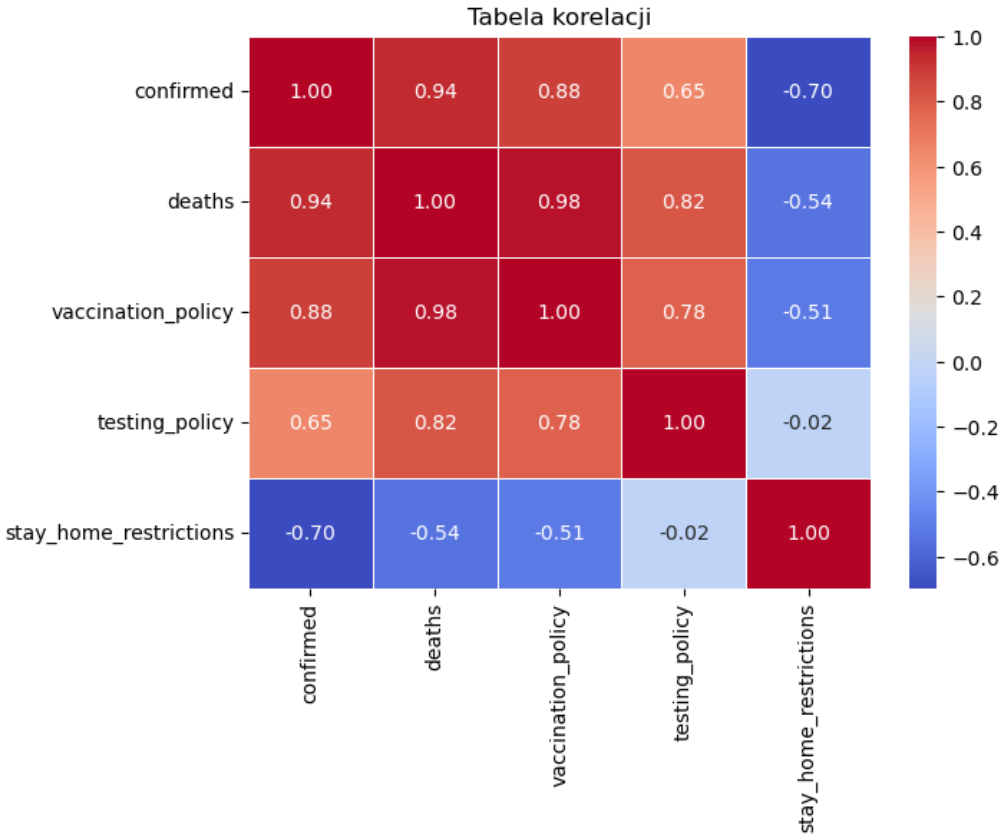
Dla najlepszej zmiennej objaśniającej (vaccination_policy) sprawdzamy czy rozkład rezyduów spełnia założenia regresji liniowej.





Dla obu zmiennych widać że, występuje zmienna wariancja reszt i histogram błędów nie jest rozkładem normalnym, model nie jest wystarczający do określenia tej zależności. Wynika z tego że nie są spełnione założenia dla regresji liniowej

Sprawdzamy współliniowość zmiennych i wybieramy te, które zostaną użyte dla regresji wielowymiarowej.



Porównujemy modele regresji liniowej do wielowymiarowej.

	objasniajaca	objasniana	mse	mae	r2
0	vaccination_policy	confirmed	10516096897410950.00	80512070.16	0.78
1	vaccination_policy	deaths	222236977436.95	429627.42	0.96
2	[vaccination_policy, testing_policy, stay_home...	confirmed	5037872956595484.00	57217264.50	0.90
3	[vaccination_policy, testing_policy, stay_home...	deaths	81726594297.10	237520.30	0.98

Wszystkie metryki dla regresji wielorakiej są lepsze, czyli model lepiej dokonuje predykcji.

Dla tych samych zmiennych dokonujemy sprawdzenia działania algorytmów SVR, drzew regresyjnych i losowego lasu regresyjnego.

Wyniki dla zmiennej "confirmed":

SVR:

Wysoki błąd kwadratowy i niski współczynnik determinacji sugerują, że model SVR nie jest odpowiedni dla tych danych lub potrzebuje dalszej optymalizacji.

Drzewo decyzyjne:

Niski błąd kwadratowy i wysoki współczynnik determinacji wskazują na bardzo dobrą dopasowanie modelu drzewa decyzyjnego do danych.

Las losowy:

Podobnie jak w przypadku drzewa decyzyjnego, niski błąd kwadratowy i wysoki współczynnik determinacji sugerują, że model lasu losowego dobrze dopasowuje się do danych.

Wyniki dla zmiennej "deaths":

SVR :

Podobnie jak dla zmiennej "confirmed", wysoki błąd kwadratowy i niski współczynnik determinacji sugerują, że model SVR nie jest odpowiedni dla tych danych.

Drzewo decyzyjne:

Bardzo niski błąd kwadratowy i wysoki współczynnik determinacji wskazują na doskonałe dopasowanie modelu drzewa decyzyjnego do danych.

Las losowy :

Podobnie jak dla zmiennej "deaths", niski błąd kwadratowy i wysoki współczynnik determinacji sugerują, że model lasu losowego dobrze dopasowuje się do danych.

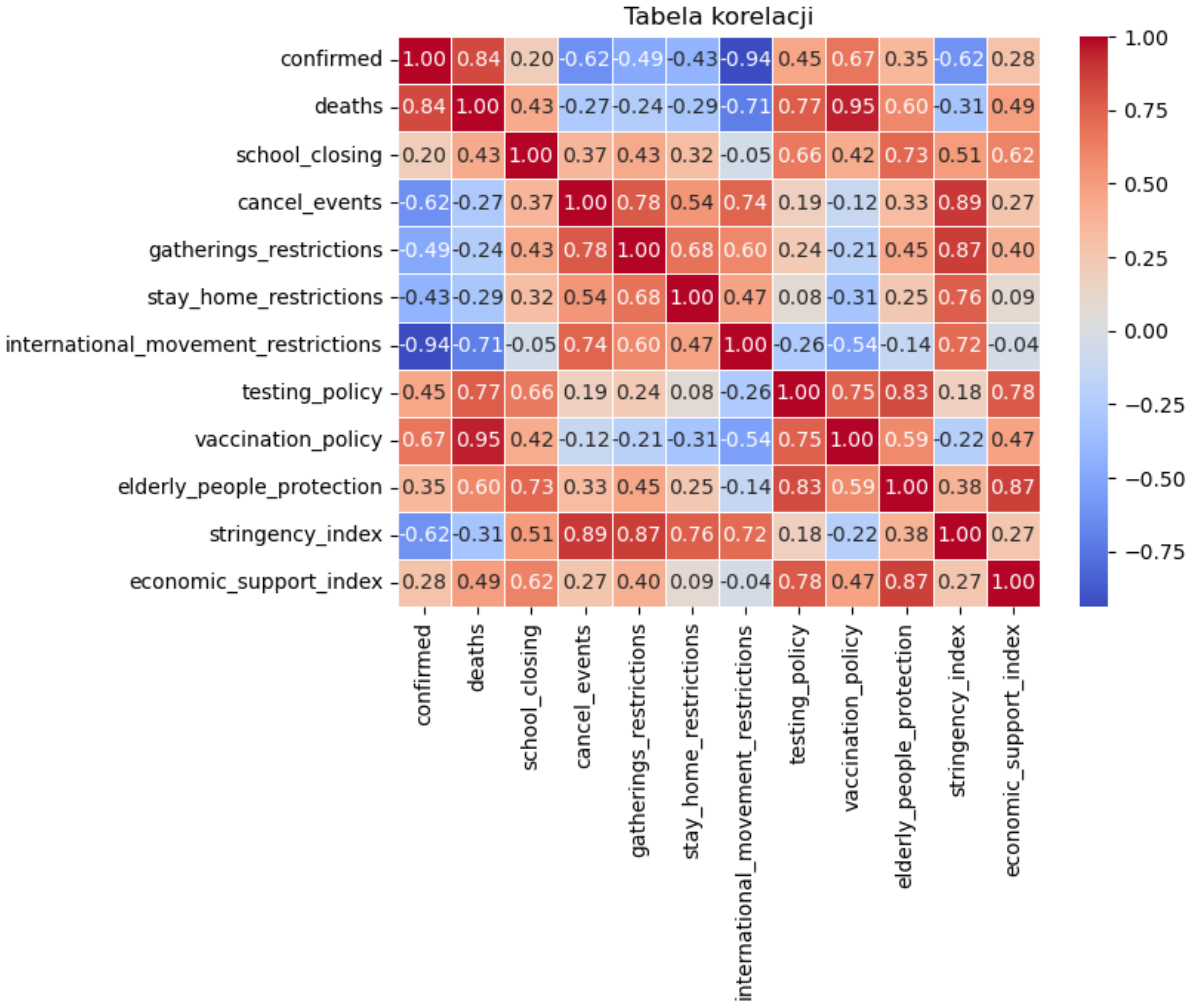
```
Results for confirmed:

SVR - Mean Squared Error: 5.1684316506636136e+16
SVR - R-squared Score: -0.068563387563646
Decision Tree - Mean Squared Error: 13309746978702.906
Decision Tree - R-squared Score: 0.9997248235193872
Random Forest - Mean Squared Error: 13209578686437.297
Random Forest - R-squared Score: 0.9997268944797276

Results for deaths:

SVR - Mean Squared Error: 5363833499803.127
SVR - R-squared Score: -0.03517214802253821
Decision Tree - Mean Squared Error: 596422405.1470097
Decision Tree - R-squared Score: 0.9998848957816667
Random Forest - Mean Squared Error: 401123074.8320096
Random Forest - R-squared Score: 0.9999225868150063
```

4. Analiza lokalna

Wybieramy kraj z najmniejszą ilością brakujących danych. Jest to Austria.

Usuwamy inne zmienne niż ilościowe i zmienną information_campaigns która ma jedną wartość.

Wybieramy najlepiej dopasowane zmienne eliminując wysokie korelacje między zmiennymi.



Wykonujemy regresję liniową od każdej zmiennej dla obu zmienny objaśnianych najlepsze dopasowanie dla zmiennej:

“Confirmed” ma zmienną international_movement_restrictions

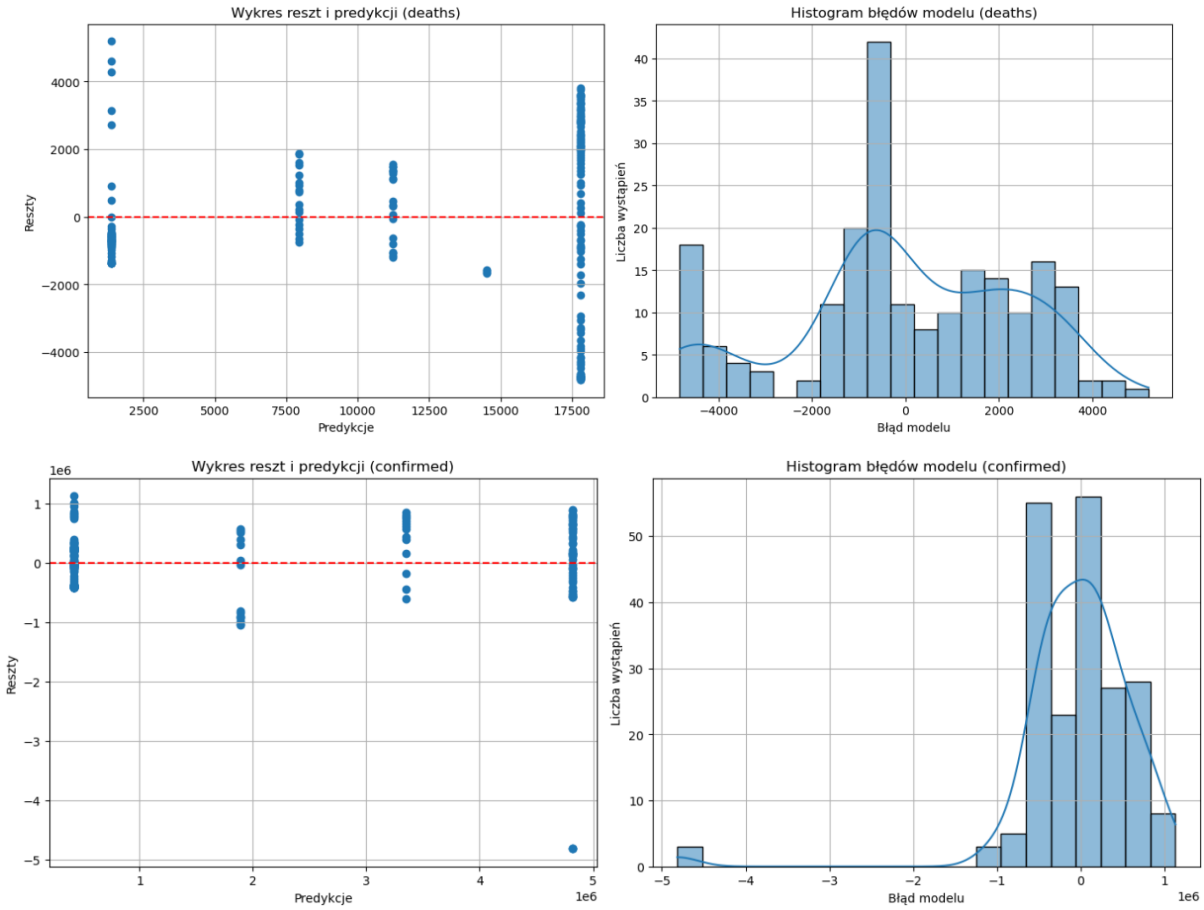
„Deaths” ma zmienną vaccination_policy

	objasniajaca	objasniana	mse	mae	r2
0	international_movement_restrictions	confirmed	544048413460.15	442437.29	0.87
1	vaccination_policy	confirmed	2176690781000.58	1215700.99	0.46
2	cancel_events	confirmed	2289353972279.49	1077815.28	0.43
3	stringency_index	confirmed	2346675784971.89	1176944.66	0.42
4	gatherings_restrictions	confirmed	2866066675418.37	1379107.55	0.29
5	testing_policy	confirmed	3194336125015.35	1543638.28	0.21
6	stay_home_restrictions	confirmed	3202502811568.95	1471710.98	0.21
7	elderly_people_protection	confirmed	3580221001764.24	1685855.15	0.11
8	economic_support_index	confirmed	3707407056369.09	1735412.55	0.08
9	school_closing	confirmed	3987659826017.90	1761593.59	0.01
10	vaccination_policy	deaths	5738320.38	1925.85	0.90
11	testing_policy	deaths	21978375.56	3963.83	0.61
12	international_movement_restrictions	deaths	28543255.83	4087.84	0.50
13	elderly_people_protection	deaths	38113467.98	5080.51	0.33
14	economic_support_index	deaths	42983824.26	5590.13	0.24
15	stringency_index	deaths	49391207.04	5928.70	0.13
16	school_closing	deaths	50335725.86	5955.45	0.11
17	cancel_events	deaths	50722445.46	6132.04	0.10
18	stay_home_restrictions	deaths	51414850.94	5874.26	0.09
19	gatherings_restrictions	deaths	51664965.58	6072.53	0.09

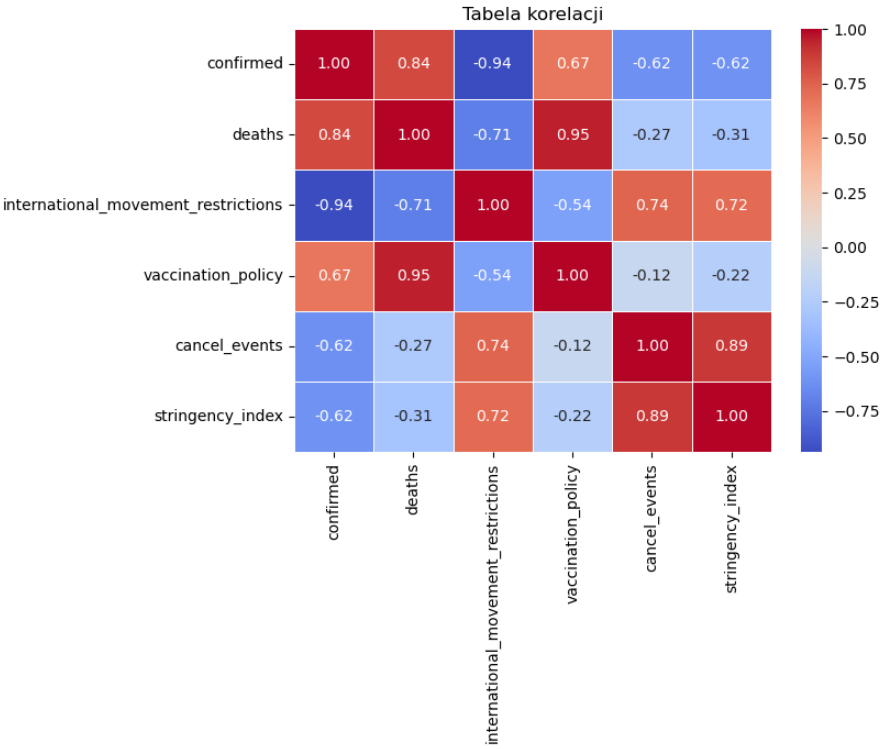
Rozkład reszt pokazuje pewną nieliniową zależność.

Widoczne są skupiska punktów, świadczy to o zmiennej wariancji.

Wniosek: nie są spełnione założenia regresji liniowej.



Wybieramy zmienne dla regresji wielowymiarowej.



Wszystkie metryki pokazują że regresja wieloraka lepiej przewiduje zmienną objaśnianą.

	objasniajaca	objasniana	mse	mae	r2
0	international_movement_restrictions	confirmed	544048413460.15	442437.29	0.87
1	vaccination_policy	deaths	5738320.38	1925.85	0.90
2	[international_movement_restrictions, vaccinat...	confirmed	363083047932.74	347634.49	0.91
3	[international_movement_restrictions, vaccinat...	deaths	2245776.06	1094.90	0.96

Dla tych samych zmiennych dokonujemy sprawdzenia działania algorytmów SVR, drzew regresyjnych i losowego lasu regresyjnego.

Wyniki dla zmiennej "confirmed":

SVR:

Wysoki błąd kwadratowy i niski współczynnik determinacji sugerują, że model SVR nie jest odpowiedni dla tych danych lub potrzebuje dalszej optymalizacji.

Drzewo decyzyjne:

Niski błąd kwadratowy i wysoki współczynnik determinacji wskazują na bardzo dobrą dopasowanie modelu drzewa decyzyjnego do danych.

Las losowy:

Podobnie jak w przypadku drzewa decyzyjnego, niski błąd kwadratowy i wysoki współczynnik determinacji sugerują, że model lasu losowego dobrze dopasowuje się do danych.

Wyniki dla zmiennej "deaths":

SVR :

Podobnie jak dla zmiennej "confirmed", wysoki błąd kwadratowy i niski współczynnik determinacji sugerują, że model SVR nie jest odpowiedni dla tych danych.

Drzewo decyzyjne:

Bardzo niski błąd kwadratowy i wysoki współczynnik determinacji wskazują na doskonałe dopasowanie modelu drzewa decyzyjnego do danych.

Las losowy :

Podobnie jak dla zmiennej "deaths", niski błąd kwadratowy i wysoki współczynnik determinacji sugerują, że model lasu losowego dobrze dopasowuje się do danych.

```
Results for confirmed:

SVR - Mean Squared Error: 5284937680717.248
SVR - R-squared Score: -0.31129068683013683
Decision Tree - Mean Squared Error: 58952039509.61732
Decision Tree - R-squared Score: 0.9853729097581124
Random Forest - Mean Squared Error: 59010090236.30433
Random Forest - R-squared Score: 0.985358506300235

Results for deaths:

SVR - Mean Squared Error: 57626279.3569331
SVR - R-squared Score: -0.017737842350896926
Decision Tree - Mean Squared Error: 189844.0704940006
Decision Tree - R-squared Score: 0.9966471634670193
Random Forest - Mean Squared Error: 191443.20381648664
Random Forest - R-squared Score: 0.9966189211700079
```

5. Podsumowanie

Wykonano szereg analiz i modeli predykcyjnych opartych na regresji w celu przewidywania liczby potwierdzonych przypadków oraz liczby zgonów związanych z COVID-19. Przeanalizowano:

1.Sprawdzenie jakości danych: zidentyfikowano brakujące dane, przekształcono niepoprawne dane.

2.Wybór zmiennych objaśniających: usunięto zmienne o niskiej korelacji z wartościami objaśnianymi, wybrano zmienne o najniższej korelacji między sobą.

3.Regresja liniowa i wieloraka: stworzono modele regresji liniowej dla pojedynczych zmiennych oraz modele regresji wielorakiej dla zestawów zmiennych.

4.Ocena modeli: sprawdzono reszty i ich rozkład dla modeli, obliczono MSE, MAE oraz R^2 dla wszystkich modeli.

Algorytmy predykcyjne oparte na regresji są dobrym rozwiązaniem tego problemu. Trzeba mieć jednak na uwadze, że model regresji liniowej nie jest najlepszym rozwiązaniem, znacznie lepiej wypada pod względem naszych metryk regresja wielomianowa.

W celu polepszenia wyników można spróbować uwzględnić tygodniowy przyrost liczby zachorowań i zgonów, użyć nieliniowych modeli.

Problem ten jest bardzo ciężki do analizy globalnej, w każdym kraju inaczej respektuje się rozporządzenia władz. W niektórych prawdopodobnie mogło ich w cale nie być. Wątpliwa jest również poprawność danych nie jesteśmy w stanie stwierdzić czy wszystkie dane są poprawne składają się a to takie czynniki jak skuteczność testów na COVID-19 czy weryfikacja śmierci spowodowanych przez tego wirusa.