COVID

Filip Przyczyna

July 13, 2024

1. Link do Colaba

https://colab.research.google.com/drive/1VEfDf9X8FJF9YAahECp6XNcAcrn9QTD9#scrollTo=BAs7wT_i2JAy

2. Dane Covid

Dane pochodzące z COVID-19 Data Hub zawierają informacje dotyczące różnych aspektów pandemii Covid-19 dla różnych krajów i jednostek administracyjnych na przestrzeni czasu. Są tu dostępne m.in. skumulowane potwierdzone przypadki danego dnia, liczba wykoannych testów, liczba śmierci, polityka epidemiologiczna, kraj i jego położenie, liczba zaszczepionych itd.

3. Wybór danych

Litwa ma najmniejszą proporcję brakujących do wszystkich danych w wyznaczonych, najważniejszych do predykcji, kolumnach. Wobec tego zostanie wybrana do lokalnej predykcji liczby śmierci i zachorowań.

4. EDA

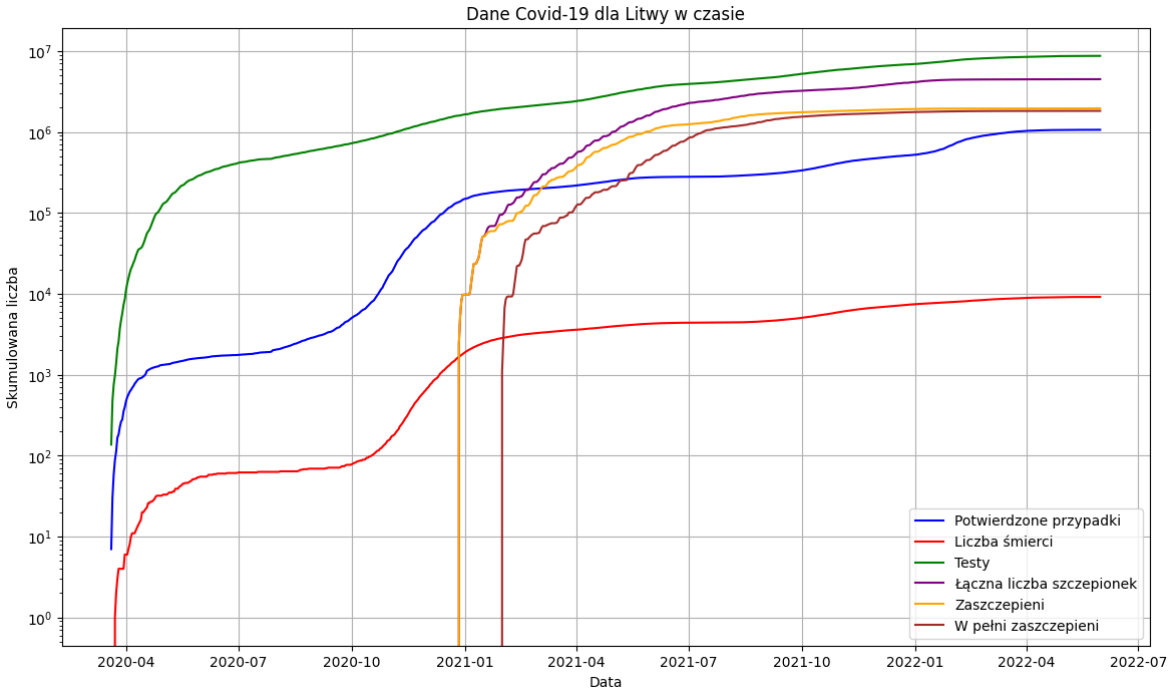

Figure 1: *Dane Covid-19 dla Litwy w czasie*

Dla wszystkich wartości na wykresie pierwsze dni mają dynamiczny wzrost. Potem trend nadal był wzrostowy, ale już mniej oczywisty, następowały również okresy wypłaszczenia. Można zauważyć, że mimo rozpoczęcia szczepień i dużej jej liczby, liczba śmierci i potwierdzonych przypadków nadal rosła. Najbardziej dynamiczne wzrosty potwierdzonych przypadków następowały w okresach październik-listopad oraz luty-marzec, czyli podczas sezonu grypowego/przejściowego między zimą a latem.

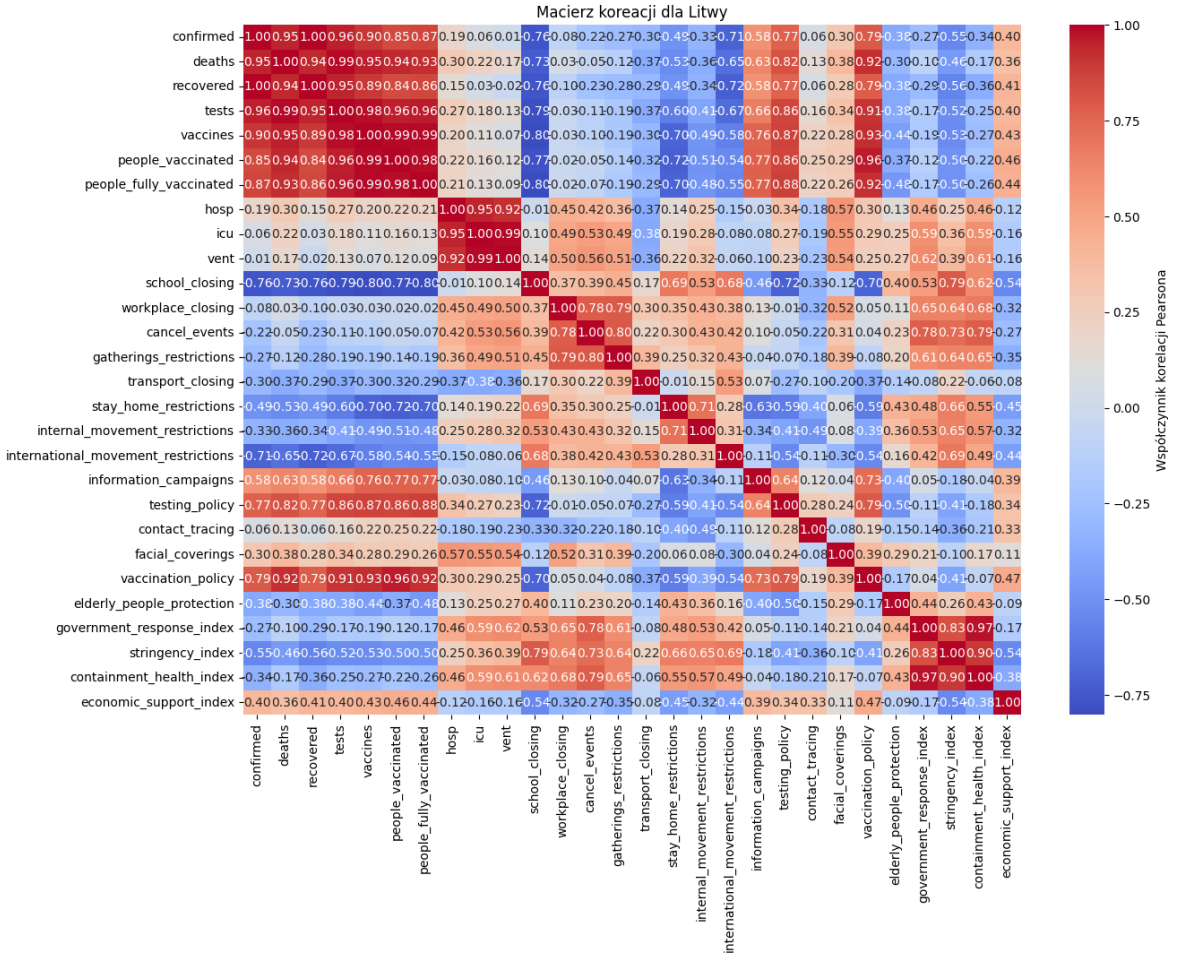

Figure 2: *Heatmap macierzy korelacji*

Heatmapa na podstawie macierzy korelacji pokazuje zależność kolejnych zmiennych od siebie. Największy współczynnik korelacji widać dla zmiennych: confirmed, deaths, recovered, tests, vaccines, people_vaccinated, people_fully_vaccinated, vaccination_policy oraz testing_policy. Najniższy zmienne te mają ze school_closing.

5. Usunięcie zbędnych kolumn

id, administrative_area_level, administrative_area_level_1, administrative_area_level_2, administrative_area_level_3, latitude, longitude, iso_alpha_3, iso_alpha_2, iso_numeric, iso_currency, key_local, key_google_mobility, key_apple_mobility, key_jhu_csse, key_nuts, key_gadm

Są to kolumny, które nie mają wpływu na analizę, np. szerokośc i długość geograficzna, numer kraju, poziom administracyjny, itp.

1 maja 2022 na Litwie skończył się stan wyjątkowy w związku z Covid-19. W danych z czerwca 2022 r. brakuje już niektórych statystyk, w związku z czym dane z tego miesiąca zostaną usunięte. Nie mają już większego znaczenia dla modelu. Po tej operacji nie ma już brakujących wartości.

6. Regresja liniowa

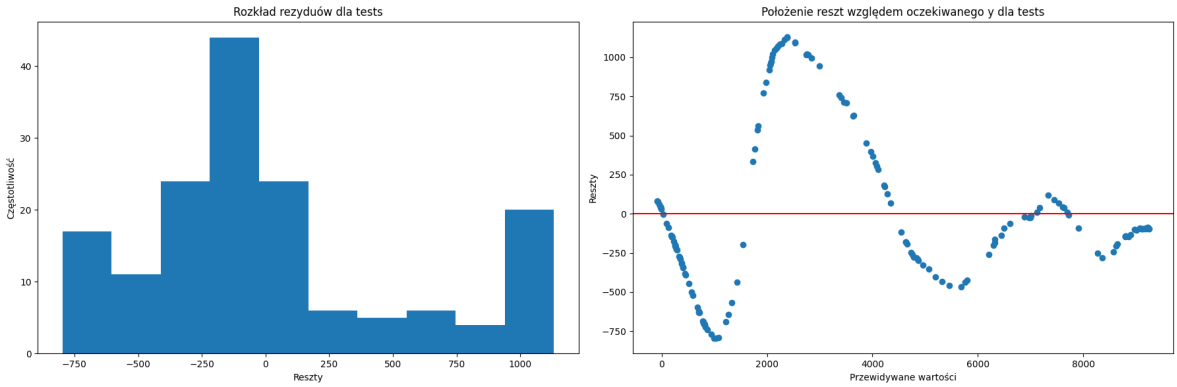

Figure 3: *Regresja liniowa predykcji liczby śmierci przy pomocy liczby testów*

W modelu regresji liniowej dla zmiennej deaths najlepsze wartości MSE, MAE i R2 ma zmienna tests (najmniejszy błąd średniokwadratowy, i średni błąd bezwzględny oraz największy współczynnik dopasowania R2 - 0.97). Również dobre, ale nie aż tak, wyniki mają zmienne recovered, people_vaccinated, people_fully_vaccinated, vaccine_policy i vaccines. Niskie współczynniki R2 otrzymała regresja liniowa dla zmiennych international_movement_restrictions, school_closing, testing_policy. Nie są to dobre modele.

	Variable	MSE	MAE	R2
0	recovered	1.122880e+06	972.624591	0.887277
1	tests	2.660797e+05	394.237249	0.973289
2	vaccines	9.968448e+05	857.923152	0.899930
3	people_vaccinated	1.315793e+06	987.051234	0.867911
4	people_fully_vaccinated	1.516137e+06	1112.791815	0.847800
5	vaccination_policy	1.472033e+06	899.870434	0.852227
6	testing_policy	3.258191e+06	1676.794235	0.672920
7	school_closing	4.904986e+06	1827.334118	0.507603
8	international_movement_restrictions	5.422047e+06	1915.267234	0.455697

Figure 4: *Porównanie róznych regresji dla liczby śmierci*

7. Założenia regresji

Histogram reszt dla zmiennej tests nie przypomina rozkładu normalnego, na skrajnych wartościach słupki są zbyt wysokie. Wobec tego nie spełnia założeń regresji liniowej.

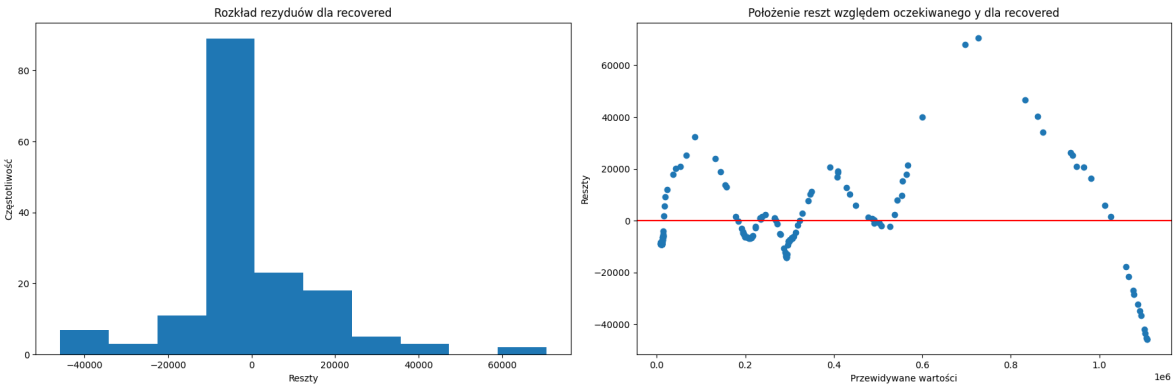

Figure 5: *Regresja liniowa predykcji liczby przypadków przy pomocy liczby wyzdrowiałych*

Histogram rezyduów dla recovered przypomina rozkład normalny. Rozkład reszt w umiarkowanym stopniu jest blisko zerowych wartości. Wobec tego można uznać, że spełnione są warunki regresji liniowej.

	Variable	MSE	MAE	R2
0	recovered	2.993812e+08	12267.186735	0.997248
1	tests	9.841200e+09	78085.858315	0.909553
2	vaccines	2.137496e+10	116055.526763	0.803550
3	people_vaccinated	3.157651e+10	133779.453297	0.709791
4	people_fully_vaccinated	2.802765e+10	135802.317269	0.742407
5	vaccination_policy	4.080283e+10	131963.655130	0.624995
6	testing_policy	4.366165e+10	174475.130233	0.598721
7	school_closing	4.780225e+10	176658.558228	0.560666
8	international_movement_restrictions	5.125111e+10	185579.124835	0.528968

Figure 6: *Porównanie róznych regresji dla liczby przypadków*

W przypadku predykcji dla confirmed, najlepszy model uzyskano dla zmiennej recovered (R2 > 0.99 i niskie w porównaniu do innych MSE i MAE). Na wykresie położenia reszt względem przewisywanych wartości również widać, że to zmienne recovered ma reszty najbliżej zera. Tests i vaccines również mają przyzwoite dopasowanie. Pozostałe zmienne nie tworzą dobrego modelu regresji.

9. Współliniowość

VIF (Variance Inflation Factor) pomaga zidentyfikować współliniowość między zmiennymi objaśniającymi. Wysokie wartości VIF (zazwyczaj > 10) sugerują, że zmienna jest silnie skorelowana z innymi zmiennymi, co może prowadzić do niestabilnych i trudnych do interpretacji wyników w modelu regresji.

Usunięcie zmiennych o wysokim VIF zmniejsza problem współliniowości, co może prowadzić do bardziej wiarygodnych i interpretowalnych modeli, choć czasem kosztem niższego R2.

	Variable	VIF
0	const	85.537939
1	tests	11.854661
2	vaccination_policy	6.407340
3	testing_policy	4.007723
4	school_closing	3.050151
5	international_movement_restrictions	2.202681

Figure 7: *VIF*

W przypadku użycia innych zmiennych VIF niższe niż 10 ma dodatkowo zmienna vaccination_policy. Te zmienne zostaną użyte do modelu regresji wielowymiarowej.

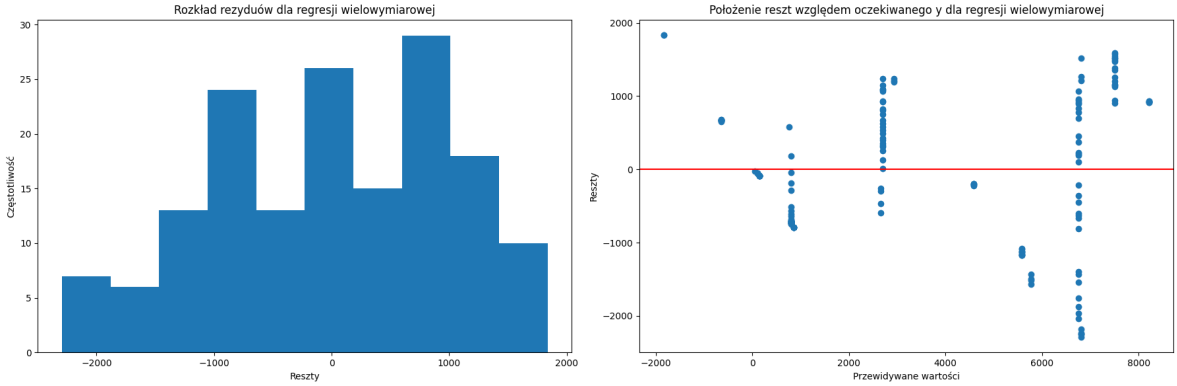

Figure 8: *Regresja wielowymiarowa dla predykcji liczby śmierci*

Model ten ma MSE i MAE rzędy wielkości niższe niż modele regresji liniowej, jednocześnie zachowując podobne dopasowanie R2. Rozkład rezyduów nie przypomina rozkładu normalnego w takim stopniu jak recovered dla predykcji confirmed, jednak jest lepszy od pozostałych regresji.

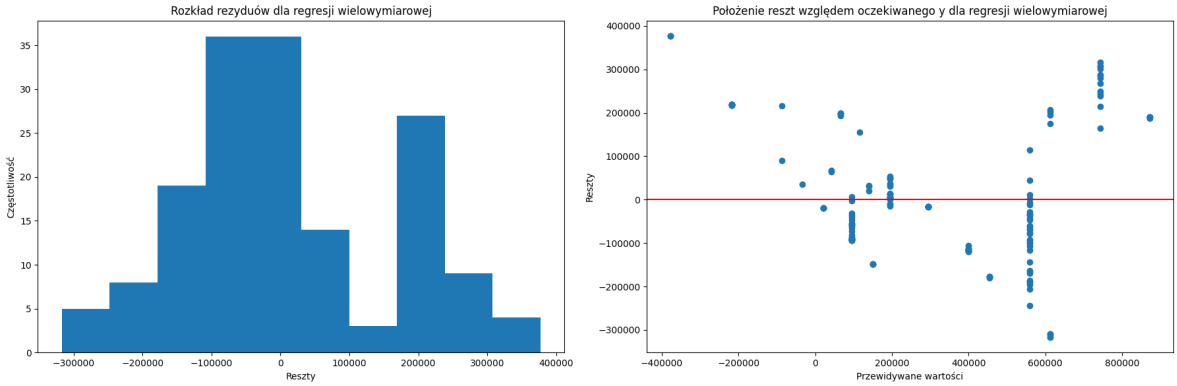

Figure 9: *Regresja wielowymiarowa dla predykcji liczby przypadków*

W tym przypadku regresje zmiennych recovered i tests dla predyckji confirmed mają niższe MSE i MAE oraz wyższe R2, wobec czego są lepszym modelem niż regresja wielowymiarowa.

11. Algorytmy - porównanie wyników

Algorithm	MSE	MAE	R2
Support Vector Regression	8963848.07	2485.60	0.10
Decision Tree Regression	370689.88	342.98	0.96
Random Forest Regression	367691.80	342.00	0.96

Table 1: Porównanie wyników różnych algorytmów regresji

Do predykcji liczby śmierci najlepsze wartości MAE, MSE oraz R2 uzyskał Random Forest Regression, nieco gorsze drzewa decyzyjne. SVR ma bardzo niskie R2.

Algorithm	MSE	MAE	R2
Support Vector Regression	113749022570.10	242379.69	-0.04
Decision Tree Regression	5233829843.55	36814.44	0.95
Random Forest Regression	5161360590.80	36652.82	0.95

Table 2: Porównanie wyników różnych algorytmów regresji

Podobnie jak dla zmiennej deaths liczba potwierdzonych przypadków ma najlepszy model dla losowego lasu regresyjnego i drzew decyzyjnych. Najgorsze dopasowanie ma SVR.

12. Podsumowanie

1. Kraj do analizy został wybrany na podstawie brakujących wartości dla kluczowych danych. 2. Dokonano EDA dla danych z Litwy, w tym macierz korelacji, pairplot, wykres przyrostu wartości w czasie. 3. Usunięto zbędne do analizy kolumny. 4. Dane do uczenia podzielone zostały w stosunku 20:80. 5. Stworzono wiele modeli regresji do predykcji liczby śmierci oraz potwierdzonych przypadków. Dla liczby śmierci najlepszy model powstał przy użyciu liczby wykonanych testów, a dla liczby potwierdzonych przypadków najlepsza okazała się liczba wyzdrowiałych. 6. Dla wybranych zmiennych sprawdzono działanie algorytmów SVR, drzew regresyjnych i losowego lasu regresyjnego, z czego słabe wyniki dał jedynie SVR. MAE i MSE modele te miały jednak wyższe od modelów regresji liniowej.

Algorytmy predykcyjne oparte na regresji nie są dobrym rozwiązaniem dla tego problemu. Wysokie wartości MSE i MAE sugerują, że model nie jest w stanie dobrze przewidywać liczby zgonów na podstawie wybranych zmiennych.

Możliwe zmiany w podejściu to: - zastosowanie innych zmiennych do predykcji, które mogą mieć większy rzeczywisty wpływ na przewidywane wartości - przekształcenie danychi i obróbka - zastosowanie innych modeli do predykcji

Podejście globalne do problemu może być utrudnione ze względu na rózne podejście do pandemii w zależności od kraju. Wpływ na wyniki mogą mieć inne aspekty, jak stan służby zdrowia. Polityka covidowa ma różne poziomy intensywności i daty zaostrzeń, co prowadzi do utrudnionej analizy. Kraje mają różne metody zbierania i raportowania danych dotyczących COVID-19, co może prowadzić do niespójności w analizach globalnych. Różnice geograficzne i społeczne oraz podejście ludzi w danym kraju do pandemii również się różni. Możliwa jest analiza i predykcja dla wybranego kraju z uwzględnieniem wszystkich przeszkód i różnic.

Na ogół liczba wykonanych testów oraz potwierdzone przypadki są wysoko skorelowane. Nie koniecznie musi tu być zależność, jednak im więcej testów jest wykonane, tym więcej zakażonych się wykrywa. Może to być przydatne przy proponowaniu polityki epidemicznej oraz do izolacji takich ludzi, jeśli zajdzie potrzeba. Z wykresów wywnioskować można, że szczepienia nie powstrzymały pandemii. Mimo wysokiej liczby zaszczepionych ludzie nadal chorowali i umierali. Większy wpływ na liczbę zachorowań mógł mieć miesiąc i sezon chorobowy.

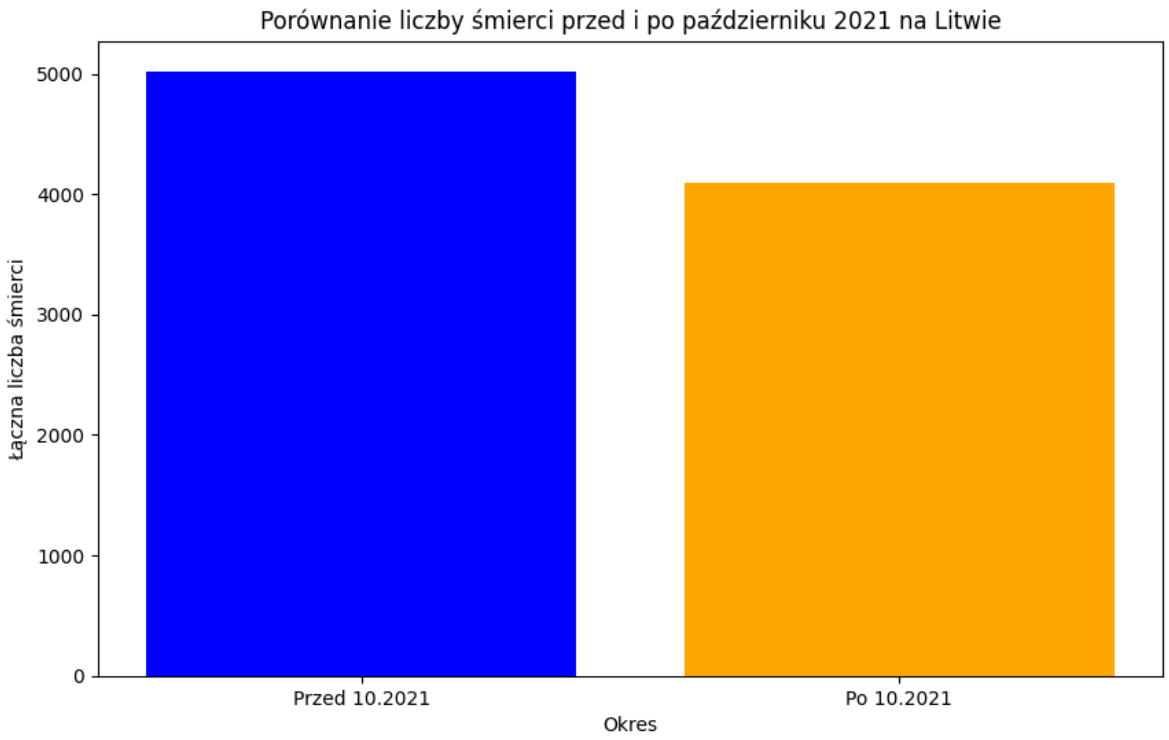

Figure 10: *Porównanie liczby śmierci przed i po październiku 2021 na Litwie*

Widać, że liczba śmierci po 1.10.2021 r., czyli od momentu gdy prawie wszyscy, którzy mieli się zaszczepić, byli już zaszczepieni, jest niewiele mniejsza niż suma wcześniej, mimo około dwukrotnie mniejszego przedziału czasowego. Nie ma jedynie pewności, czy wszystkie śmierci spowodowane były Covidem.

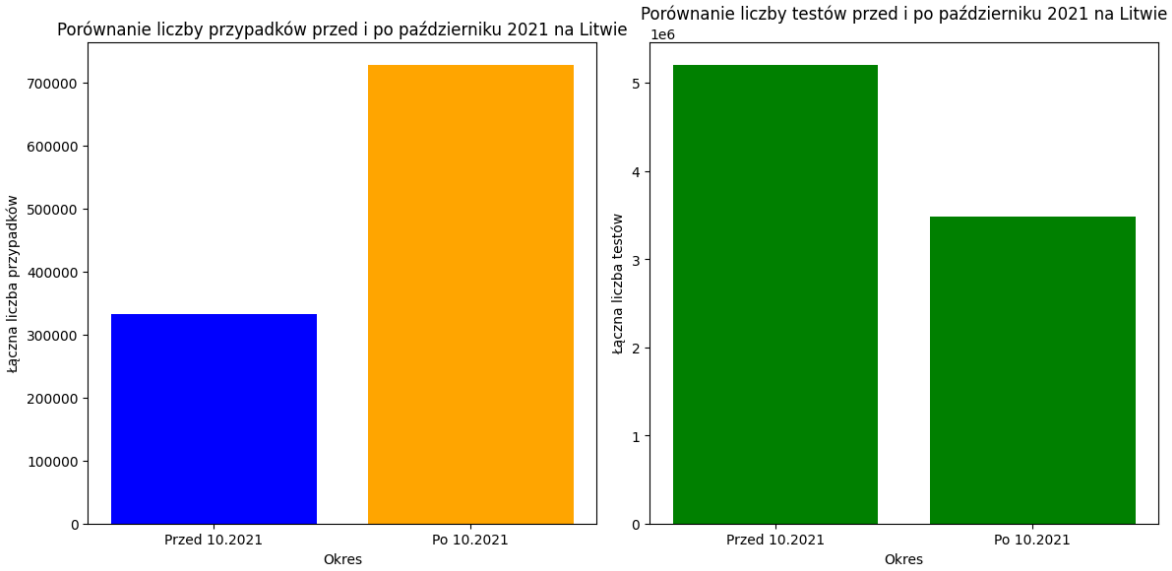

Figure 11: *Porównanie liczby przypadków i testów przed i po październiku 2021 na Litwie*

W analogiczny okresie czasu zauważyć można, że mimo mniejszej liczby wykonanych testów po zaszczepieniu społeczeństwa, liczba przypadków była znacznie wyższa niż wcześniej. Wobec tego szczepienia mogły spowodować mniejszą zaraźliwość. Mogło to nastąpić z innych powodów. Żeby to jednoznacznie stwierdzić, należałoby dokonać dokładniejszej analizy.