Jakub Muzyka i Łukasz Serafin

Pakiety statystyczne raport nr 1

18 grudnia 2022

1. Wstęp

Do wykonania sprawozdania wykorzystaliśmy dane z popularnej strony internetowej:
https://www.kaggle.com/datasets/caesarmario/our-world-in-data-covid19-dataset/, które dotyczą wirusa COVID-19. Dane pochodzą z okresu od 01.01.2020 roku do 08.12.2022 roku. Zawierają one 239947 wierszy i 67 kolumn. Udowodnimy bądź zaprzeczymy tezie, że **"Im więcej szczepień tym mniej zgonów oraz hospitalizacji"**. Celem naszej analizy jest przedstawienie różnych aspektów dotyczących szczepionek na COVID-19 oraz zobrazowanie danych w celu udowodnienia lub obalenia postawionego przez nas pytania badawczego.

1.1. Opis zmiennych wykorzystanych w sprawozdaniu

W naszej pracy znajdują się zarówno zmienne kategoryczne oraz ciągłe. Poniżej przedstawiamy nazwy kolumn, których użyliśmy do wykonania raportu:

1. **Continent** (zmienna kategoryczna) — jak sama nazwa mówi, kolumna ta posiada nazwy kontynentów, na których rozwijał się wirus (na przykład: Europa, Azja, ...).
2. **Location** (zmienna kategoryczna) — przechowuje nazwy państw, w których pojawiały się zakażenia (na przykład: Polska, Argentyna, Rosja, ...).
3. **Date** (zmienna ciągła, określona w formacie dzień/miesiąc/rok) — daty od 01.01.2020 roku do 08.12.2022 roku, przedstawiające okres, z którego zostały zebrane dane (na przykład: 26.01.2020, 11.11.2021, ...).
4. **Total_cases** (zmienna kategoryczna) — liczba wszystkich potwierdzonych przypadków zarażenia się wirusem COVID-19 (na przykład: 42, 1103, ...).
5. **New_cases** (zmienna kategoryczna) — nowe, potwierdzone przypadki zarażenia się wirusem COVID-19 (na przykład: 1, 17, ...).
6. **Total_deaths** (zmienna kategoryczna) — liczba wszystkich zgonów, które są wynikiem zarażenia się wirusem COVID-19 (na przykład: 2, 571, ...).

7. **New_deaths** (zmienna kategoryczna) — nowe przypadki zgonów w wyniku zarażenia (na przykład: 5, 12, ...).
8. **Icu_patients** (zmienna kategoryczna) — Liczba pacjentów zarażonych COVID-19, którzy przebywają na oddziale intensywnej terapii w danym dniu (na przykład: 5, 10, ...).
9. **Hosp_patients** (zmienna kategoryczna) — Liczba hospitalizowanych pacjentów w danym dniu (na przykład: 3, 7, ...).
10. **Total_vaccinations** (zmienna kategoryczna) — Liczba wszystkich zarejestrowanych i przyjętych szczepionek (na przykład: 2, 6500, ...).
11. **people_vaccinated** (zmienna kategoryczna) — Liczba osób, które przyjęły przynajmniej jedną dawkę szczepionki (na przykład: 4, 2.5, ...).
12. **People_fully_vaccinated** (zmienna kategoryczna) — liczba osób, które otrzymały wszystkie dawki określone w protokole szczepienia (na przykład: 75, 421, ...).
13. **Total_boosters** (zmienna kategoryczna) — Całkowita liczba podanych dawek przypominających szczepionki przeciwko COVID-19 (dawki podane poza liczbą określoną w protokole szczepienia) (na przykład: 1, 17, 55, ...).

Dodatkowo w wybranej przez nas bazie danych znajduje się wiele kolumn, które odrzuciliśmy, ponieważ nie jesteśmy w stanie z ich pomocą odpowiedzieć na postawione przez nas pytanie badawcze. Do przykładowych kolumn odrzuconych należą:

1. **Iso_code** (zmienna kategoryczna) — kolumna ta przechowuje trzyliterowy kod państwa (na przykład: AFG - Afganistan).
2. **New_cases** (zmienna kategoryczna) - nowe, potwierdzone przypadki zachorowań na COVID-19.
3. **New_cases_smoothed i New_deaths_smoothed** (zmienne kategoryczne) — obie kolumny posiadają dane wygładzone co siódmy dzień dla danych adekwatnych do nazwy kolumny.
4. **New_deaths** (zmienna kategoryczna) — przechowuje nowe, potwierdzone przypadki śmierci w wyniku zakażenia wirusem.
5. **excess_mortality_cumulative_absolute** (zmienna kategoryczna, określona w procentach) — procentowa różnica pomiędzy odnotowaną tygodniową lub miesięczną liczbą zgonów w latach 2020–2021 a prognozowaną liczbą zgonów w tym samym okresie na podstawie poprzednich lat.
6. **Total_cases_per_milion, New_cases_per_milion, New_cases_smoothed_per _million, Total_deaths_per_million, New_deaths_per_million, New_deaths_ smoothed_per_million, Icu_patients_per_million, Hosp_patients_per_million** (zmienne kategoryczne) — kolumny zawierające dane zawarte w nazwie kolumn (na przykład: per_milion - dane podzielone przez milion ludzi).
7. **Reproduction_rate** (zmienna kategoryczna) — Oszacowanie w czasie rzeczywistym efektywnego współczynnika reprodukcji (R) COVID-19.

1.2. Dane wybrane do analizy

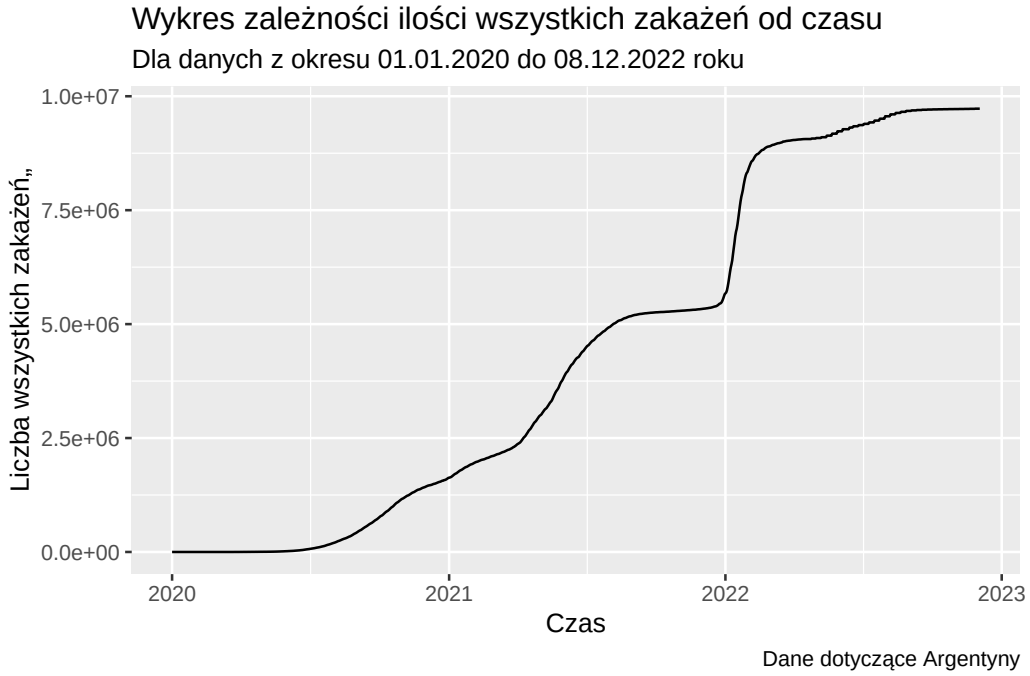
Spośród wszystkich dostępnych danych, ze względu na duże braki danych w zależności od kraju w jakim były one zbierane postanowiliśmy wyłonić 5 krajów, na których przeprowadzimy analizę. Są to odpowiednio: Argentyna, Stany Zjednoczone, Niemcy, Australia i Korea Południowa. Wybierając te kraje mieliśmy także na uwadze na jakim kontynencie znajdują się te kraje, by analiza obejmowała jak najwięcej obszarów, a nie tylko na przykład Europę. Afryka nie znajduje się w tym zestawieniu, ponieważ żaden kraj afrykański nie dysponował danymi potrzebnymi do analizy.

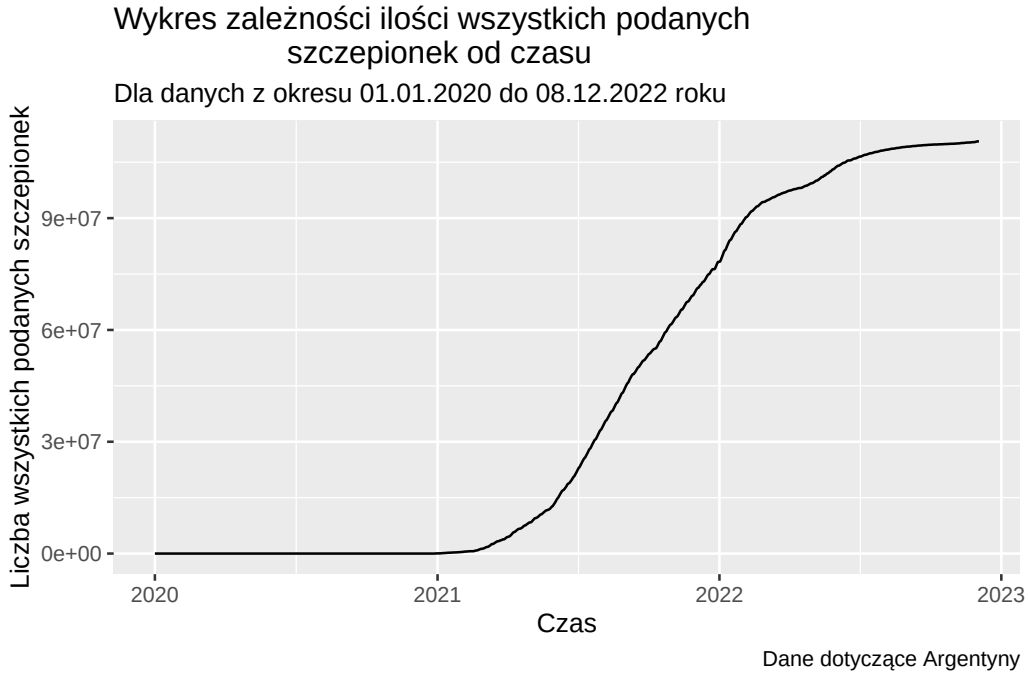
1.3. Obsługa braków wartości

Braki wartości w wybranych przez nas danych wystepowały najczęściej w pierwszych 150 wierszach, czyli tych które dotyczyły początku pandemii (okres od stycznia 2020 do lipca 2020). W przypadku kolumny total_cases braki spowodowane są faktem iż do wystąpienia pierwszej wartości nie odnotowano żadnych przypadków zakażeń, dlatego też nie prowadzono wtedy ich spisu. Braki do pierwszej wartości wypełniamy zerami. Takim samym sposobem wypełniamy 'pierwsze' braki w innych kolumnach. Natomiast pojedynczo występujące braki danych w dalszych wierszach wypełniamy wartościami poprzednimi. Na przykład: Dla kolumny icu_patients w podzbiorze dotyczącym Argentyny dla dat (16.02.2021 - 20.02.2021) wartości wynoszą kolejno: 3581,3573, NA, 3608, 3605. Po obsłużeniu braku wartości wyglądają następująco: 3581, 3573, 3573, 3608, 3605. Dla reszty kolumn postępujemy w sposób identyczny. Przy wyborze sposobu obsługi błędów kierowaliśmy się tym, aby wypełnienia braków miały logiczne podstawy oraz nie zaburzały kształtu danych. Dlatego też nie wybraliśmy metody wypełniania średnią wartością. Dla wcześniej przytoczonego przykładu średnia kolumny icu_patients wynosi po zaokrągleniu do całkowitej wartości 1891. Wypełnienie tą wartością i otrzymanie ciągu wartości: 3581, 3573, 1891, 3608, 3605 jedynie utrudniłoby analizę.

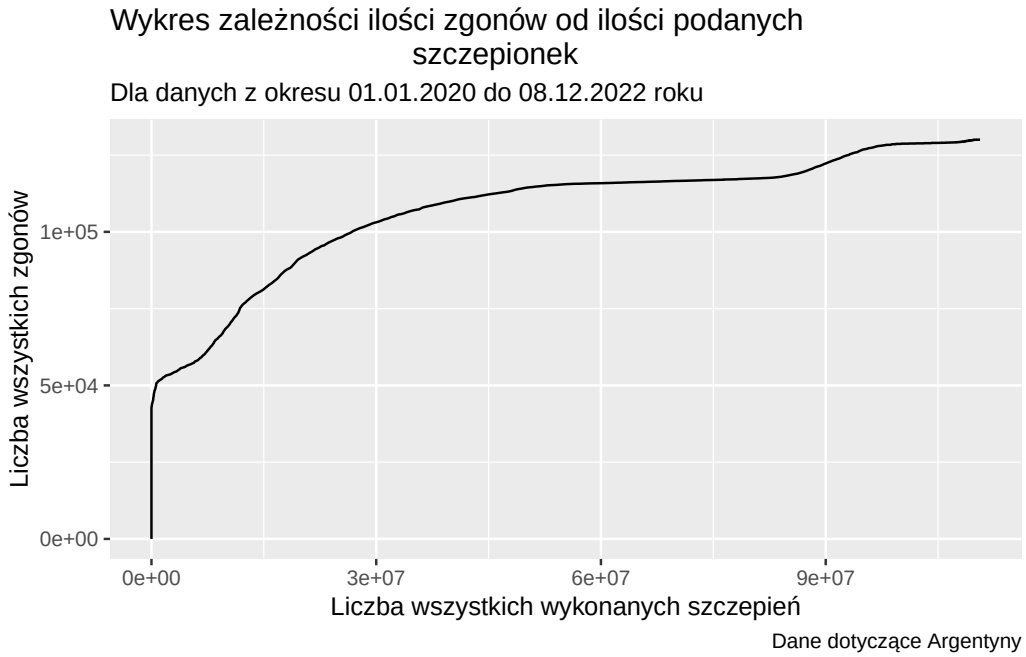
2. Analiza danych na wykresie

Na początek przedstawiliśmy dane dotyczące wszystkich zakażeń wirusem COVID-19 oraz liczbę szczepionek przyjętych przez ludzi w czasie pandemii w Argentynie.

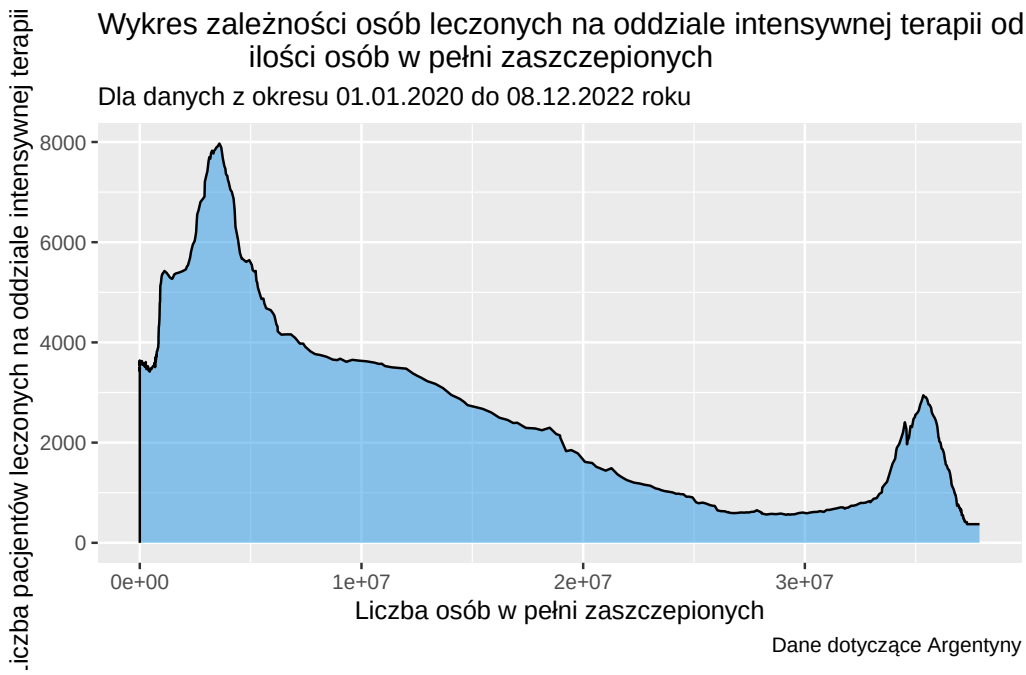




Widzimy, że liczba szczepionek rośnie niemal wykładniczo w okresie pomiędzy 2020 a 2022 rokiem, co powoduje mniej gwałtowny wzrost liczby wszystkich zakażeń w kraju. Tak przedstawione dane przeważają na korzyść postawionej przez nas tezy. Za nim jednak przyjmiemy lub odrzucimy tezę, weźmy pod uwagę, wartości w innych kolumnach.

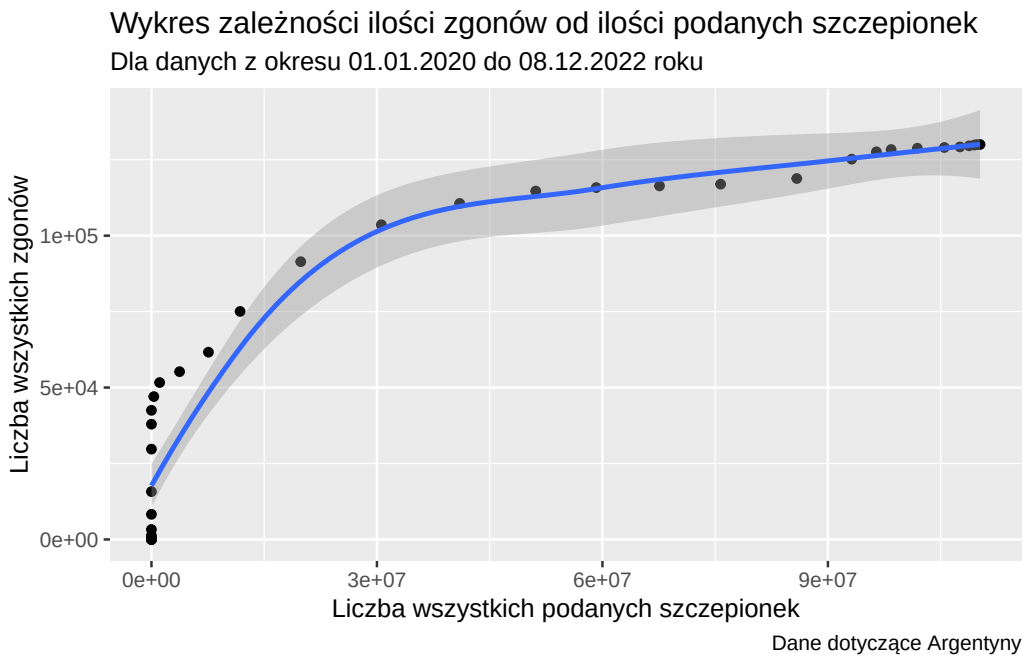


Jesteśmy w stanie zauważyć wyhamowanie wzrostu liczby zgonów w zależności od zwiększającej się ilości wykonywanych szczepień. Informacje te są potwierdzeniem danych z poprzednich wykresów, które przedstawiają nam realny wpływ szczepień na ilość nowych zakażeń, które mogą końcowo prowadzić do śmierci.



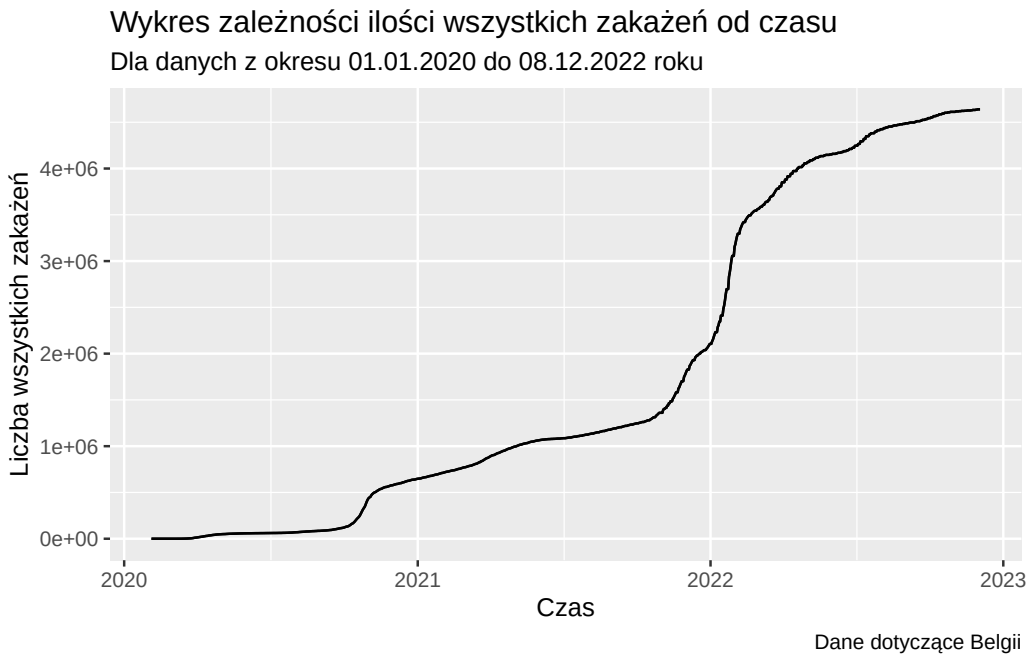
Do pewnego momentu,obserwujemy spadek osób hospitalizowanych wraz ze zwiększającą się ilością osób w pełni zaszczepionych. Na końcowym odcinku danych sytuacja ulega zmianie i liczba osób przybywających w szpitalach

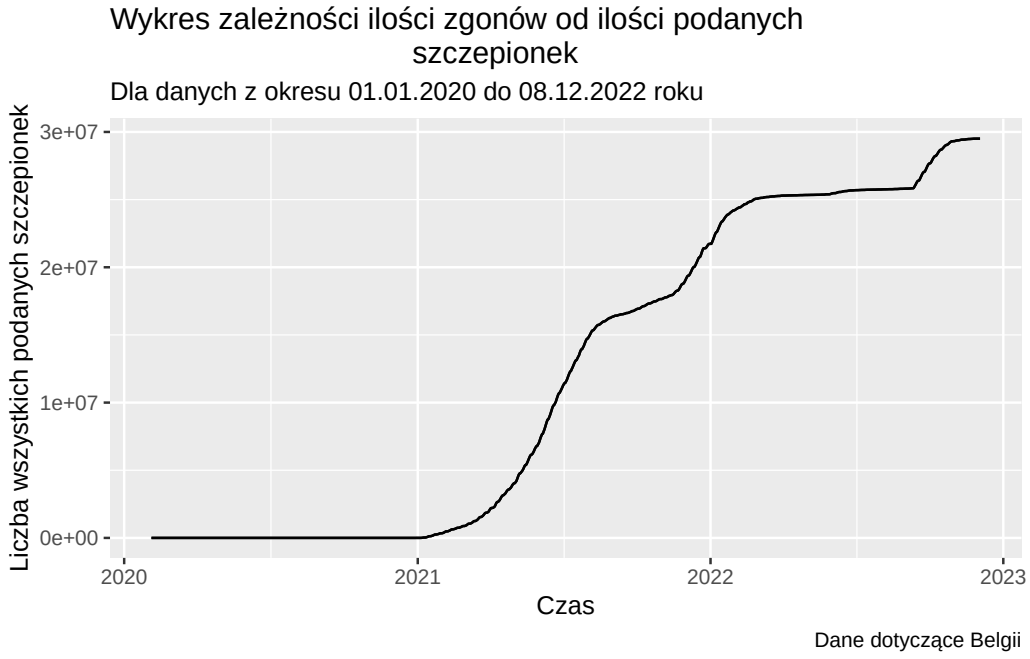
ponownie wzrasta. Zakładamy, że sytuacja ta ma związek ze znacznym, a nawet całkowitym zniesieniem restrykcji w większości krajów na świecie, co dotyczyło również Argentyny.



Dodatkowo, przedstawiamy prostą najlepszego dopasowania do danych wyżej opisanych, w celu jeszcze lepszego zobrazowania wypłaszczenia przyrostu zgonów w stosunku do ilości wykonanych szczepień.

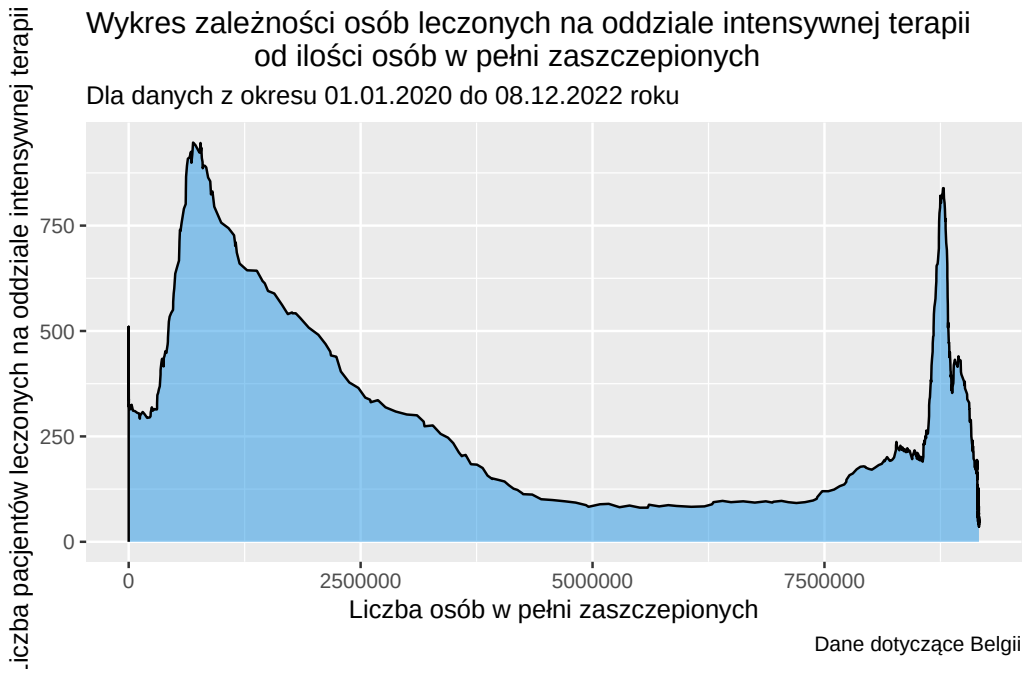
Teraz przedstawimy dane zebrane w Belgii:





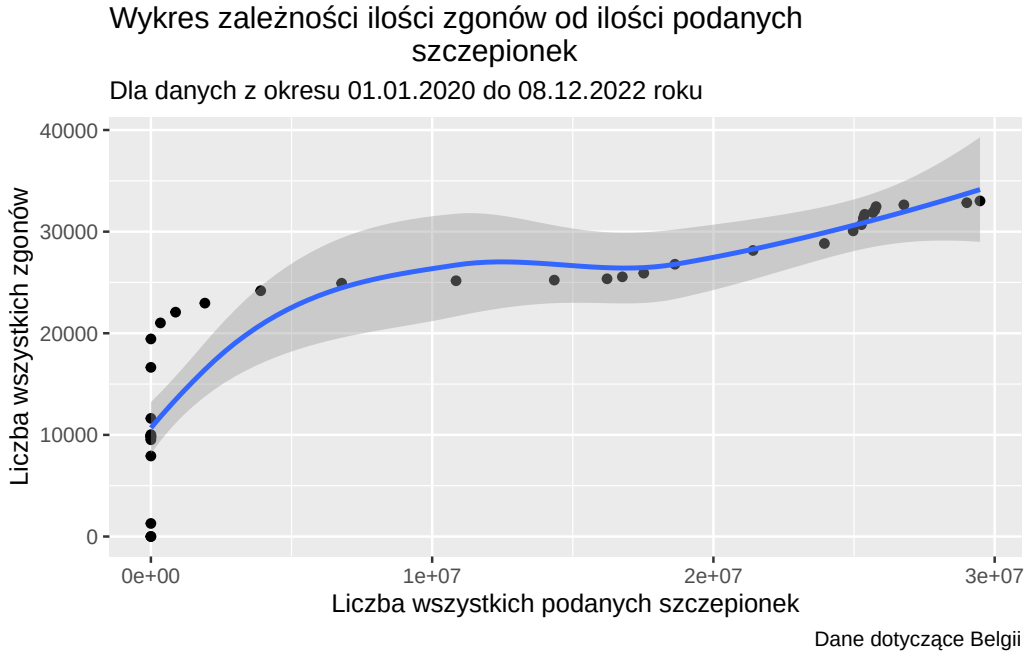
Dla pierwszych dwóch wykresów zauważamy sytuację bardzo podobna do tej zachodzącej w Argentynie. Liczba wszystkich podanych szczepionek w roku 2021 zwiększała się niemal eksponencjalnie ,na przykładzie Belgii. Pomimo dużej ilości szczepień szczyt okres największej ilości nowych zachorowań przypada na przełom roku 2021 i 2022, podobnie jak w Argentynie. Wyciągnięte stąd wnioski działają na niekorzyść naszej tezy, gdyż nie widać zmniejszonej zachorowalności przy dużej ilości szczepień.

Na potwierdzenie tezy działają jednak dane dotyczące zgonów. Po rozpoczęciu szczepień ilość zgonów drastycznie wyhamowała, nabierając rozpędu jedynie pod koniec pandemii, przy ostatniej fali zakażeń.



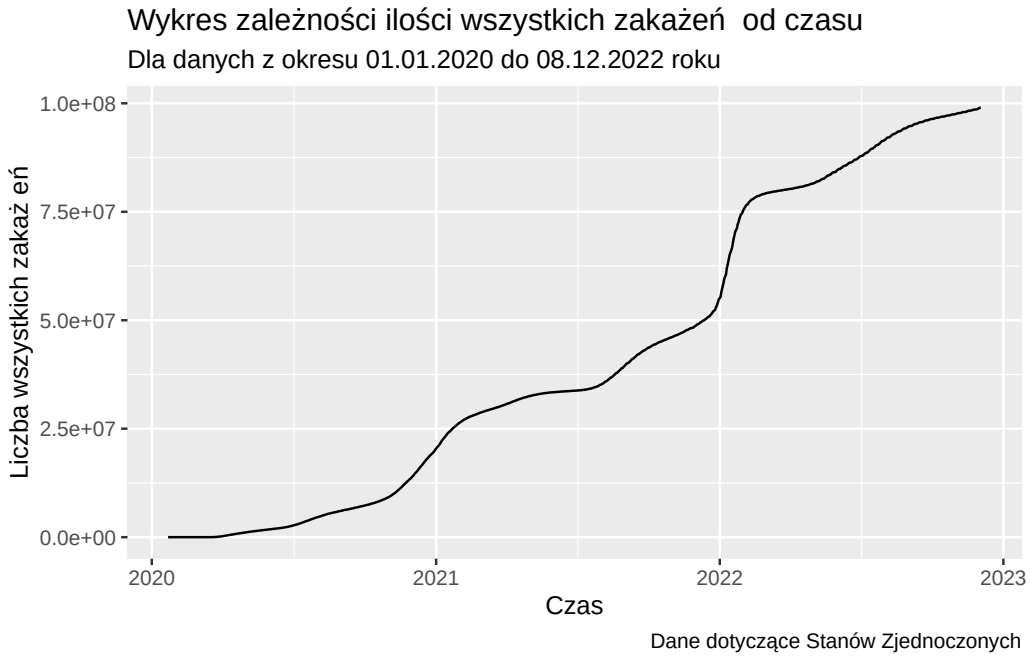
Przy kolejnym wykresie warto zwrócić uwagę na wartości osi y, czyli na ilość osób hospitalizowanych. W porównaniu do Argentyny około 10 razy mniej osób leczyło się w szpitalu, ale również około 4 razy mniej osób zostało w pełni zaszczepionych. Jest to oczywiście normalna różnica ze względów terytorialnych — Belgia jest dużo mniejszym państwem, o dużo mniejszej po-

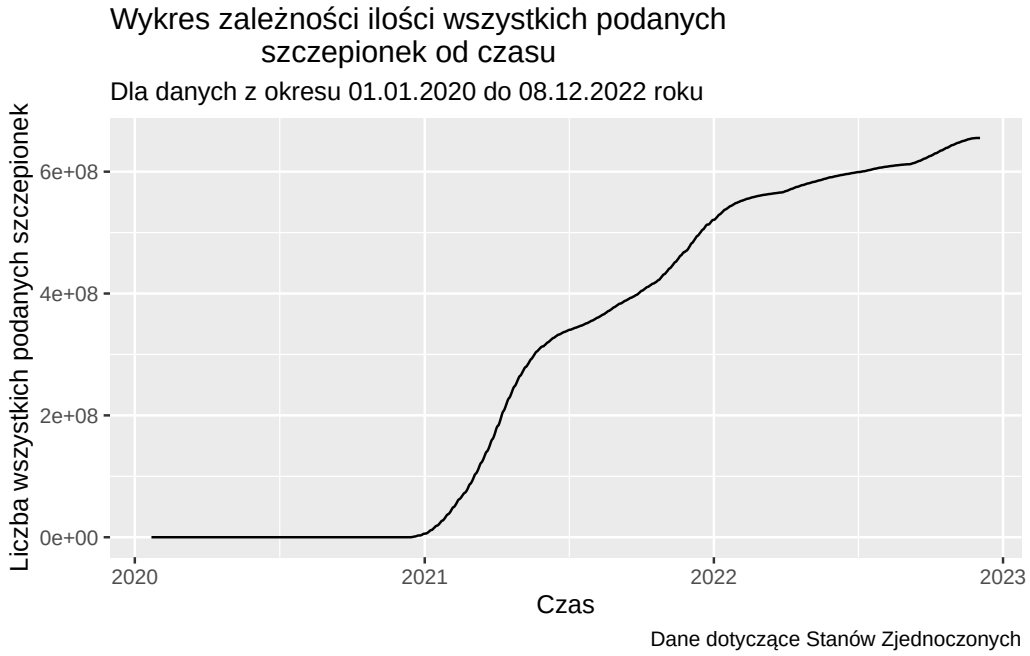
wierzchni. Nie mniej jednak możemy zauważyć, że Belgowie gorzej poradzili sobie z powrotem do "normalnego" życia, bez restrykcji, niż Argentyńczycy. Może to być spowodowane surowszymi ograniczeniami w danym państwie w czasie pandemii, ale są to tylko nasze przypuszczenia.

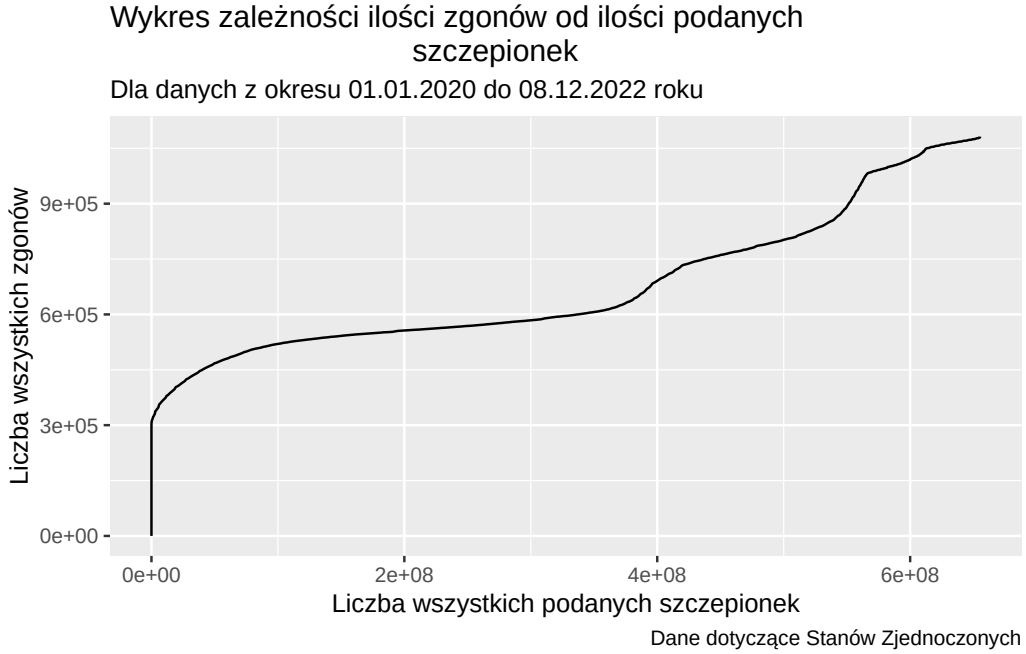


Potwierdziliśmy wnioski wyciągnięte z dwóch pierwszych wykresów, mianowicie im więcej szczepionek, tym mniej zgonów,czyli krzywa ilości zgonów jest bardziej wygładzona.

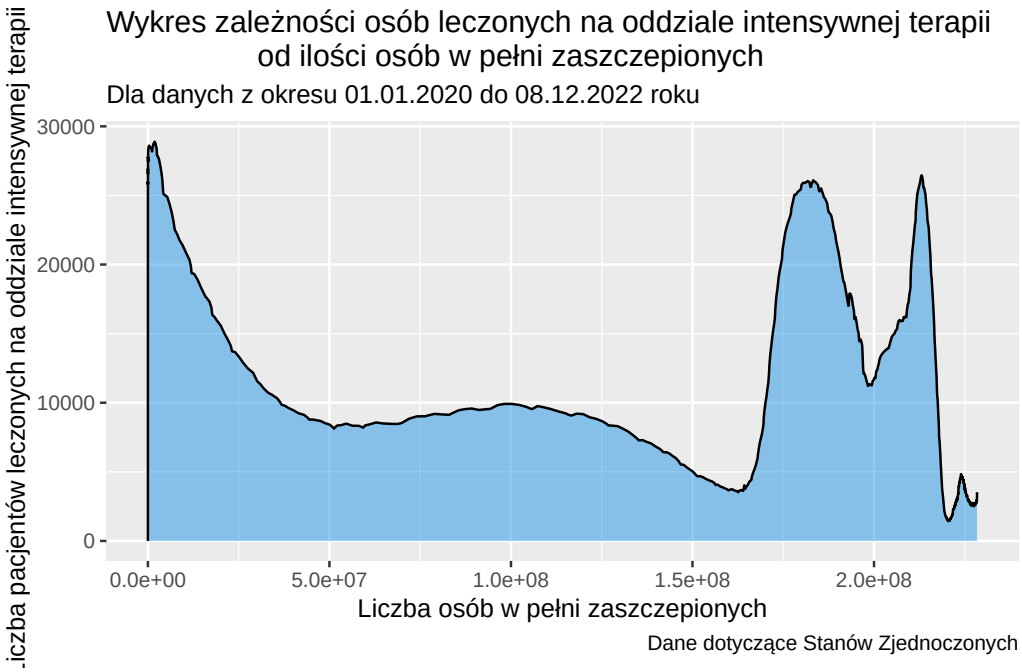
Dane dotyczące Stanów Zjednoczonych.

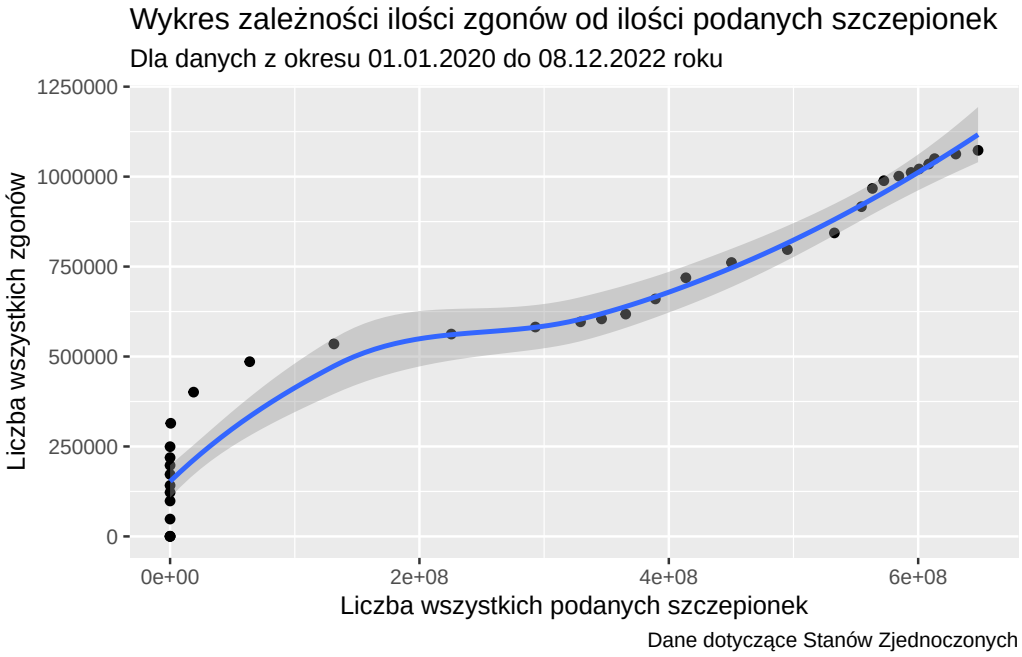






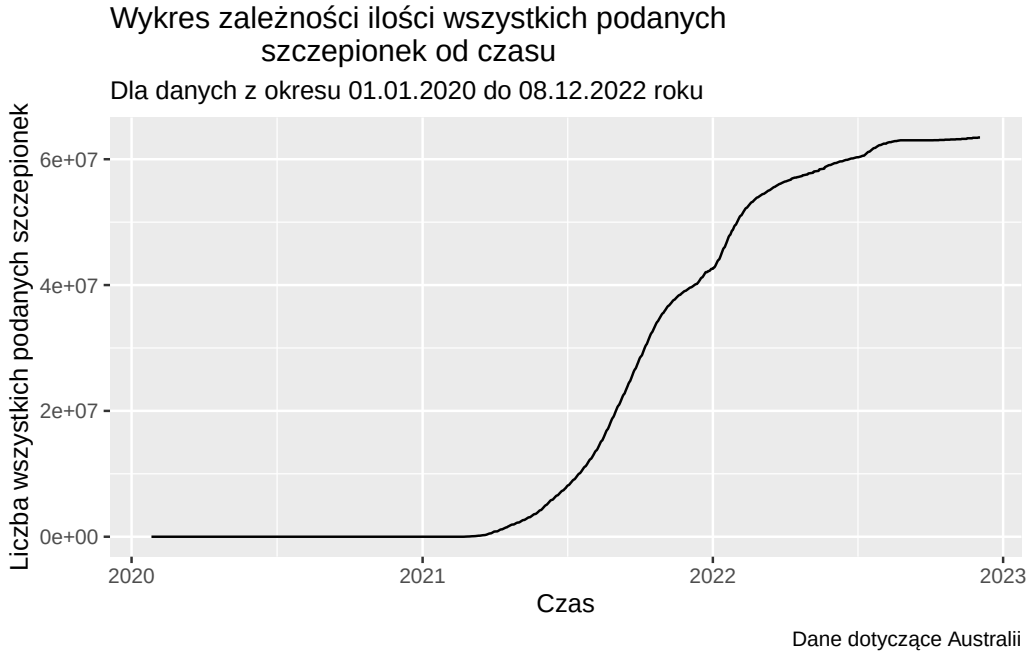
Sytuacja prezentuje się podobnie jak w poprzednich dwóch krajach, mianowicie — wraz ze wzrostem wykonanych szczepień wzrost liczby zgonów jest coraz mniejszy, lecz w tym przypadku tylko do trzeciego kwartału 2021 roku, potem ponownie zwiększa się, choć nie tak gwałtownie jak na początku pandemii. Rozbieżność pojawia się w wykresie zależności wszystkich zakażeń od czasu. Nie mamy już postaci nawet zbliżonej do eksponencjalnej, natomiast krzywa zaczyna przypominać funkcję liniową. Możemy na tej podstawie wnioskować, że Amerykanie nie odczuli za bardzo spadku liczby zakażeń po pojawieniu się szczepionek na COVID-19. Przypuszczamy, że sytuacja ta spowodowana była lekceważącym podejściem Amerykanów do restrykcji podczas pandemii oraz gęstego zaludnienia tego kraju, co ułatwiało roznoszenie wirusa.



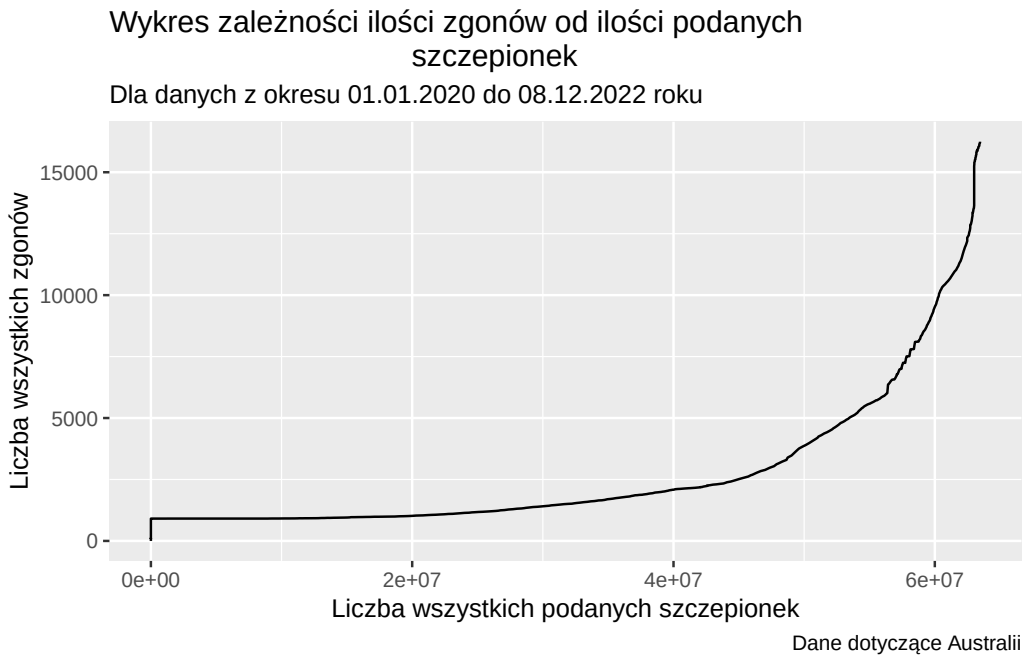


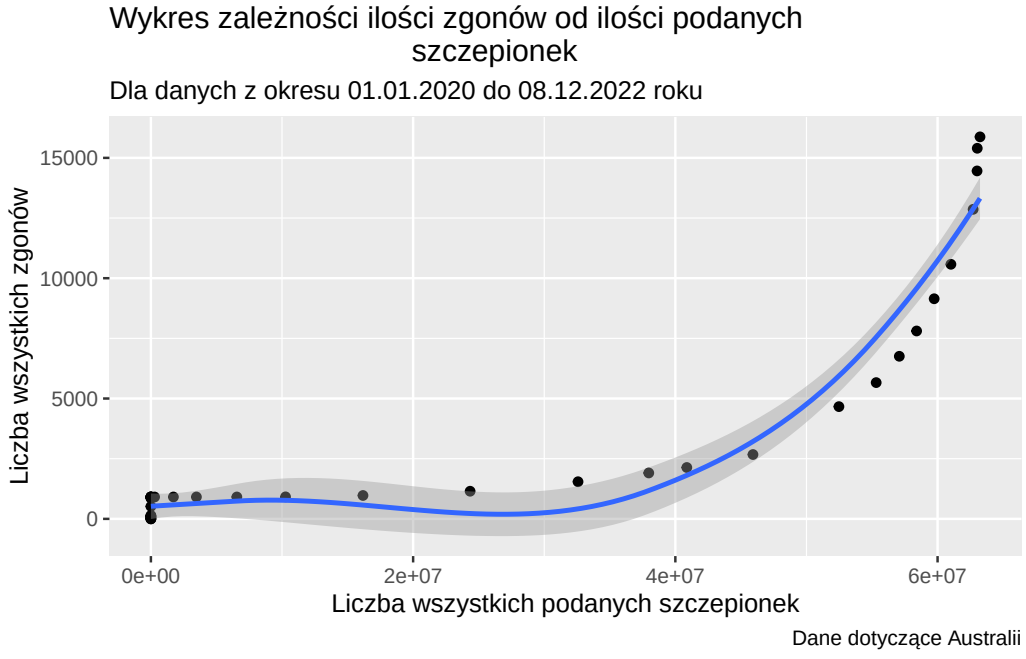
Powyższe wykresy potwierdzają nasze przypuszczenia odnośnie do lekceważącego podejścia Amerykanów, ale prawdopodobnie po zniesieniu obowiązujących ograniczeń, ponieważ na nowo wzrosła liczba hospitalizowanych pacjentów.

Dane dotyczące Australii:

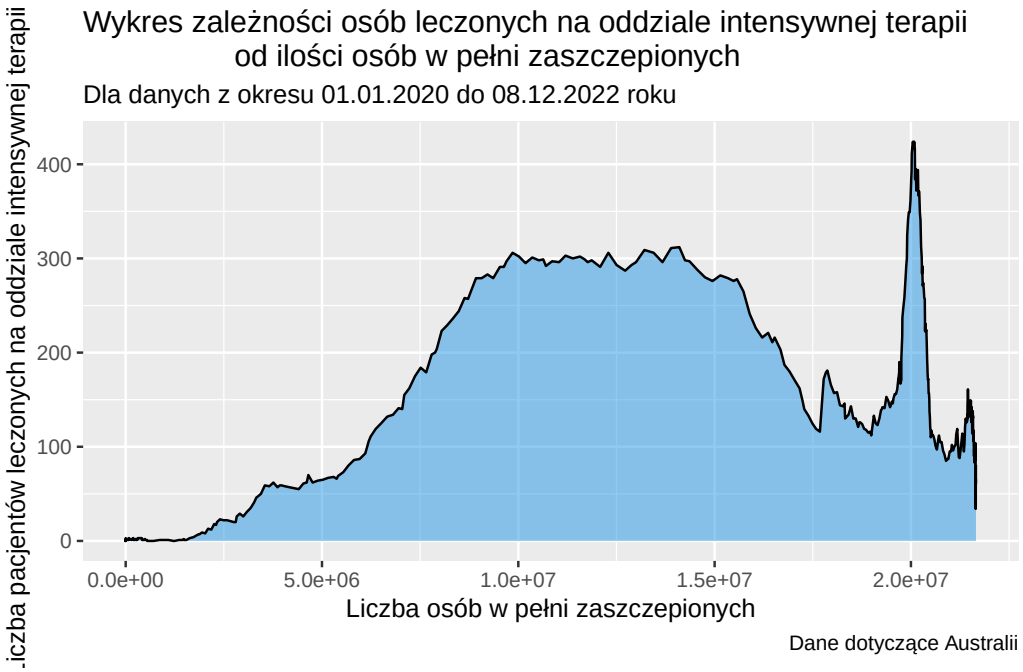


W przypadku tego kraju mamy dość nietypową sytuację, ponieważ mimo wprowadzenia szczepionek, a nawet w okresie przed możemy zauważyć bardzo spokojne wzrosty liczby zakażeń. Przypuszczamy, że było to spowodowane surowymi restrykcjami, po których zniesieniu, od 2022 roku możemy zauważyć gwałtowny wzrost zarażeń.





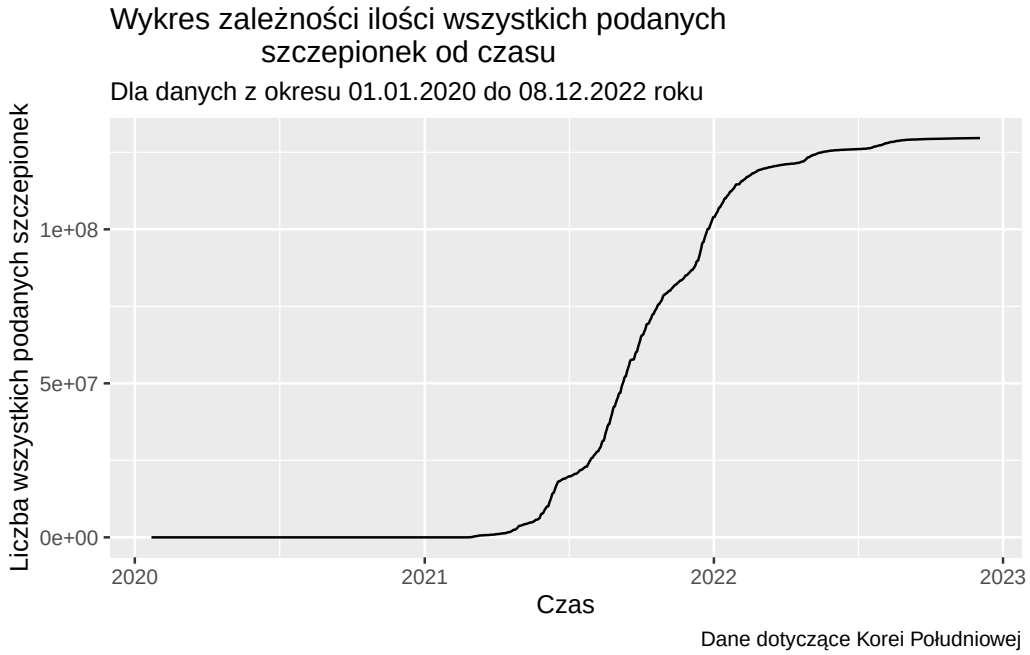
Widzimy, że razem ze wzrostem liczby szczepionek rośnie wykładniczo liczba wszystkich zgonów, co może być prawdą, w kontekście naszych wcześniejszych założeń o surowych obostrzeniach, które zapobiegały zarówno zakażeniom jak i zgonom. Przypadek Australii ciężko jest ustosunkować do naszej tezy, ponieważ nawet wysoki poziom szczepień nie zmniejszył ilości nowych zgonów ani zachorowań, lecz w przypadku braku porównania z okresem bez szczepionek oraz restrykcji, nie jesteśmy w stanie stwierdzić czy szczepienia obniżyły potencjalną ilość zgonów i hospitalizacji czy też nie.



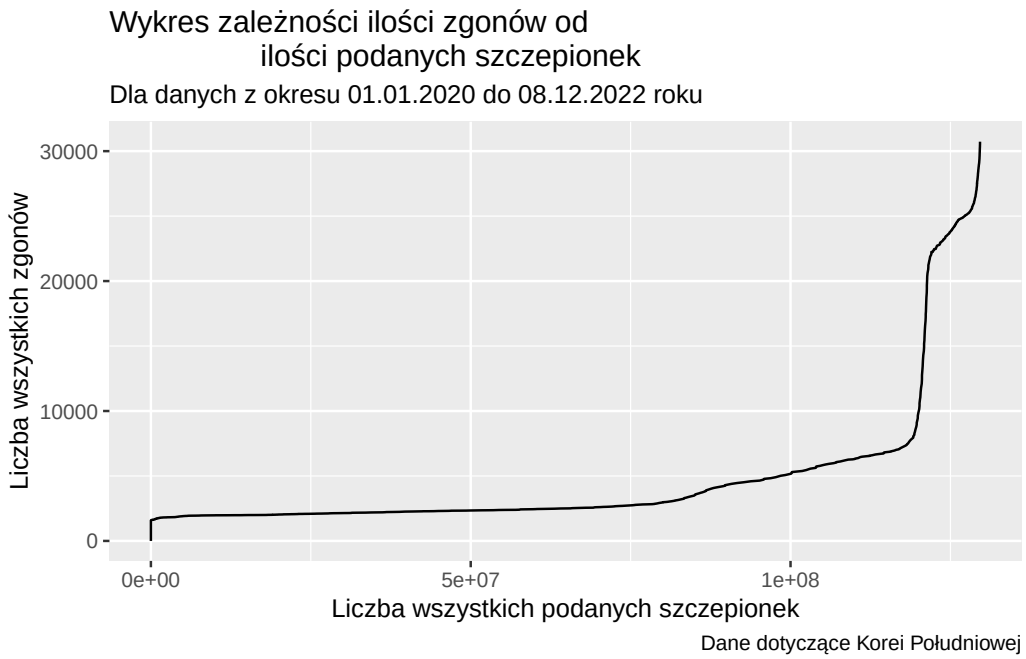
Widzimy, że szczyt ilości osób hospitalizowanych w związku z pandemią jest największy prawdopodobnie już po zniesieniu restrykcji. Maksymalna ilość wynosi 424, co w porównaniu do ponad 26 milionów mieszkańców jest to bardzo małą ilością. Mogłoby to świadczyć o skuteczności szczepionek w walce z wirusem, z drugiem jednak strony szczyt hospitalizacji przypada na okres, gdy większość populacjii jest już wyszczepiona. Są to sprzeczne wnioski i biorąc pod uwagę politykę walki z pandemią podjętą przez rząd Australii, polegającą na rygorystycznych restrykcjach, nie można więc jednoznacznie stwierdzić jaki wpływ mają szczepienia na ilość hospitalizacji oraz zgonów spowodowanych przez COVID-19. Trzeba mieć jednak na uwadze to, że liczby zgonów oraz osób leczonych na oddziałach intensywnej terapii są małe w porównaniu do ogółu populacji, co można interpretować jako argument na popracie tezy.

Ostatni omówiony przez nas zestaw danych dotyczy Korei Południowej:

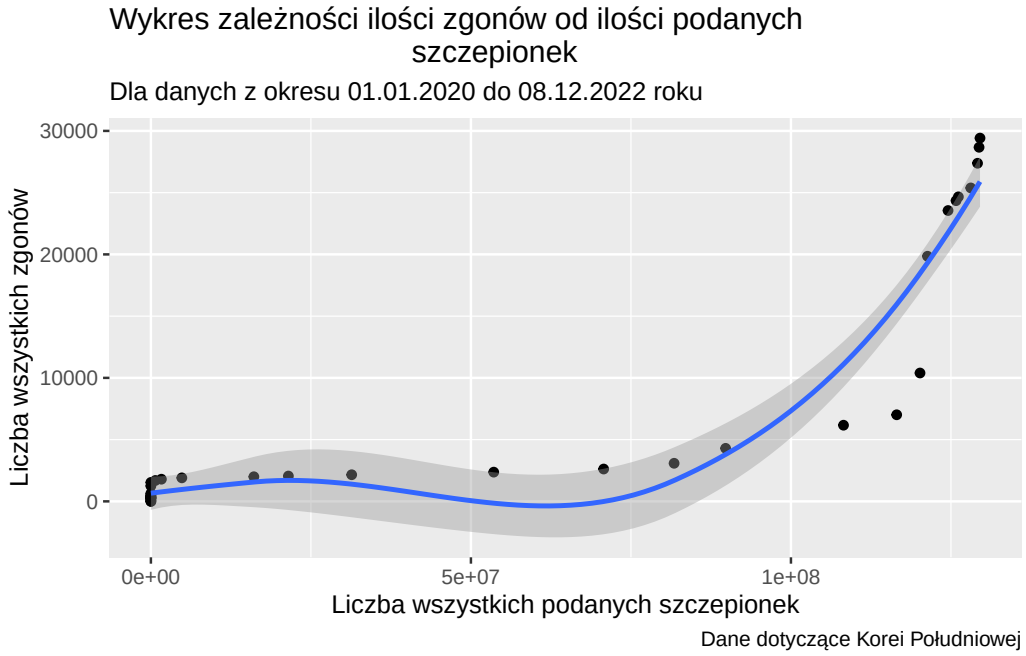
W Korei pierwsze większe ilości zakażeń są mocno odsunięte w czasie przez działania rządu, co odczytujemy z powyższego wykresu.



Szczepienia zaczęły się w czasie bardzo zbliżonym do reszty krajów, tj. I kwartał 2021 roku. Do połowy 2022 w pełni wyszczepiono ponad 45 milionów ludzi, co stanowi około 81 procent 55-milionowej populacji.



Pomimo tak dużej liczby szczepień po zniesieniu restrykcji liczba zgonów wzrosła wykładniczo, co świetnie widać na powyższym wykresie.



Z samego nachylenia linii trendu moglibyśmy wywnioskować, że im większa ilość szczepień, tym większa ilość zgonów, co jest tezą odwrotną do pierwotnie przyjętej.

Podobnie zachowała się ilość pacjentów przyjmowanych na oddziały intensywnej terapii, która w ostatnim czasie bije historyczne rekordy. Trzeba jednak brać pod uwagę strategię walki z pandemią jaką przyjął rząd Korei. Bardzo surowe restrykcje, podobnie jak w Australii, przełożyły się na bardzo niską ilość zgonów oraz hospitalizacji. Po ich poluźnieniu, drastycznie wzrosła, co można argumentować bagatelizacją wirusa. Pomimo dużych wzrostów, zarówno ilość pacjentów oddziałów intensywnej terapii jak i łączna ilość zgonów jest wyraźnie mniejsza w porównaniu do Argentyny, gdzie ilość zgonów była 4 razy większa. Z tego możemy wnioskować o skuteczności restrykcji, ale co ważniejsze, skuteczności szczepionek przy zmniejszaniu ilości zgonów i hospitalizacji.

3. Wnioski i podsumowanie

Po dogłębnej analizie danych dla poszczególnych krajów na świecie jesteśmy w stanie zweryfikować postawioną przez nas tezę. Nie została ona w pełni potwierdzona, bliższe prawdy będzie stwierdzenie, że **Im więcej szczepień tym mniej zgonów, ale nie hospitalizacji**. Na każdym podzbiorze danych dotyczącym danego kraju wykazaliśmy, że ilość szczepionek owszem, na początku pandemii ma wpływ na pomniejszenie się ilości chorych w szpitalach, natomiast w całym rozpatrywanym okresie danych widzieliśmy pod koniec wzrosty, które w niektórych krajach osiągały bardzo wysokie wartości Nie możemy przez to powiedzieć, że ilość szczepień chroni ludzi przed leczeniem się w szpitalu po zakażeniu wirusem, a więc nie zmniejsza ilości hospitalizacji. Ważnym czynnikiem mającym wpływ na liczbę hospitalizacji była taktyka walki z pandemią wybrana przez rząd danego kraju. Niezależnie jednak od wprowadzonych restrykcji, przy wysokim poziomi wyszczepienia

społeczeństwa zmiejszała się śmiertelność wirusa, co potwierdzałoby drugą część postawionej przez nas tezy.