Sprawozdanie nr 1

Klaudia Janicka 262268 i Natalia Iwańska 262270

2022-12-15

1. Wstęp

Niniejszy raport stanowi analizę danych rzeczywistych dotyczących trzęsień ziemi o skali wydarzenia powyżej 6 na przestrzeni lat 1900-2013. Dane pochodzą ze strony kaggle.com. Zgodnie z informacjami zawartymi na podanej stronie źródłem danych jest United States Geological Survey.

Celem naszej analizy jest odpowiedzenie na pytanie jakie regiony są najbardziej narażone, zbadanie jakie czynniki mają wpływ na występowanie trzęsień ziemi oraz, czy poszczególne cechy trzęsienia są ze sobą powiązane.

1.1 Opis zmiennych

Typy danych występujących w poszczególnych kolumnach:

	typ
X	integer
Date	character
Time	character
latitude	numeric
longitude	numeric
depth	numeric
mag	numeric
magType	character
nst	numeric
net	character
id	character
updated	character
place	character
type	character

Typy zawarte w tabeli **??** oznaczają odpowiednio:

integer - typ całkowity, który przyjmuje wartości całkowite,

numeric - typ zmiennoprzecinkowy, który przyjmuje wartości ułamkowe,

character - typ znakowy, przechowuje łańcuchy tekstowe.

Do analizy wykorzystujemy dane zawarte w kolumnach:

- *place* - tekstowy opis regionu geograficznego w pobliżu zdarzenia;

- *latitude* - szerokość geograficzna podana w stopniach, przyjmuje wartości z przedziału $[-90, 90]$, gdzie wartości ujemne oznaczają południowe szerokości;
- *longitude* - długość geograficzna podana w stopniach, przyjmuje wartości z przedziału $[-180, 180]$, gdzie wartości ujemne dotyczą zachodnich długości;
- *mag* - skala zdarzenia;
- *nst* - liczba stacji sejsmicznych użytych do określenia lokalizacji trzęsienia;
- *time* - czas wystapienia trzęsienia;
- *depth* - głębokość zdarzenia w kilometrach;
- *date* - data zajścia zdarzenia;

gdzie jako zdarzenie rozumiemy wystąpienie trzęsienia ziemi.

Jako zmienne kategoryczne uznajemy zmienną *place*, która po transformacji wskazuje na kraj, w którym zdarzenie miało miejsce oraz *date*, która po transformacji oznacza miesiąc zajścia zdarzenia. Pozostałe zmienne to zmiennne ciągłe. (Tak myślę, ale może być inaczej;;; brzmi sensownie)

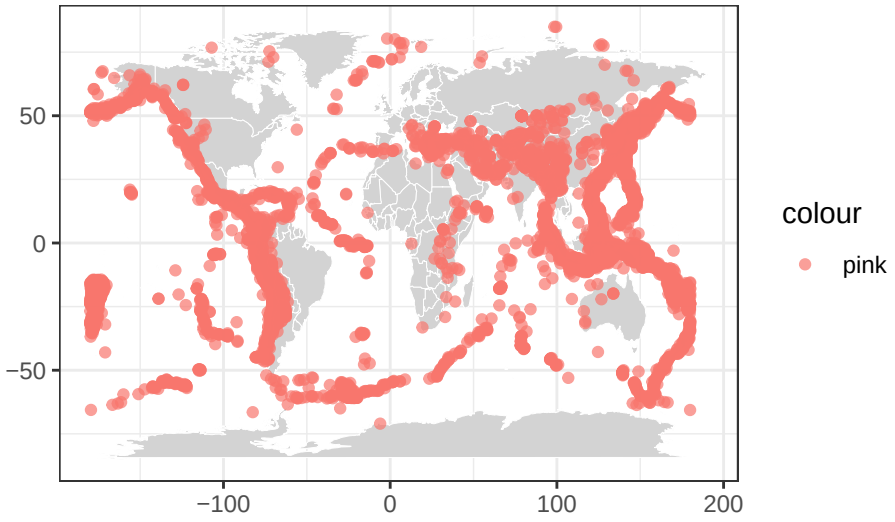
1.2 Obsługa błędów

Wiersze z brakami danych usuwamy za pomocą funkcji drop_na() z biblioteki dplyr.

2. Analiza danych

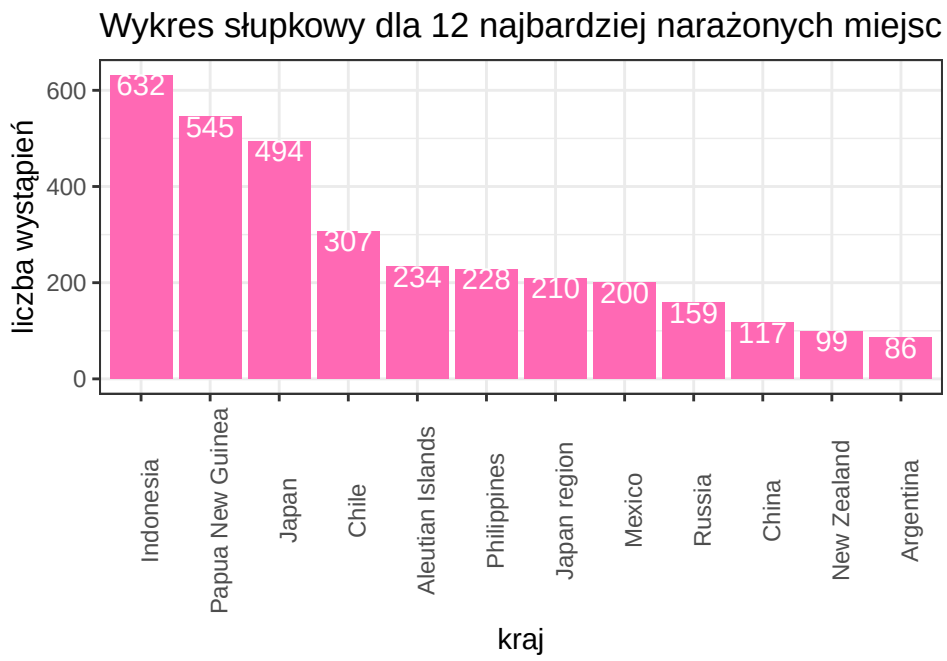
2.1 Miejsca najbardziej narażone na trzęsienia ziemii

W celu ustalenia państw najbardziej narażonych na wystąpienie zdarzenia posłużymy się analizą graficzną.



Wykres 1: Mapa świata z naniesionymi miejscami trzęsień ziemi

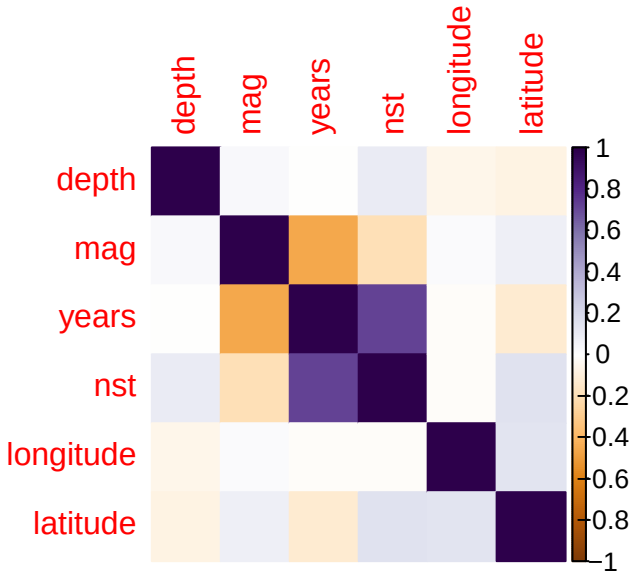
Z powyższego wykresu 1 i posiadanej wiedzy z zakresu geografii, jesteśmy w stanie stwierdzić, że najbardziej narażone są między innymi Chile, Japonia i inne kraje azjatyckie.



Wykres 2: Wykres słupkowy dla 12 najbardziej narażonych miejsc

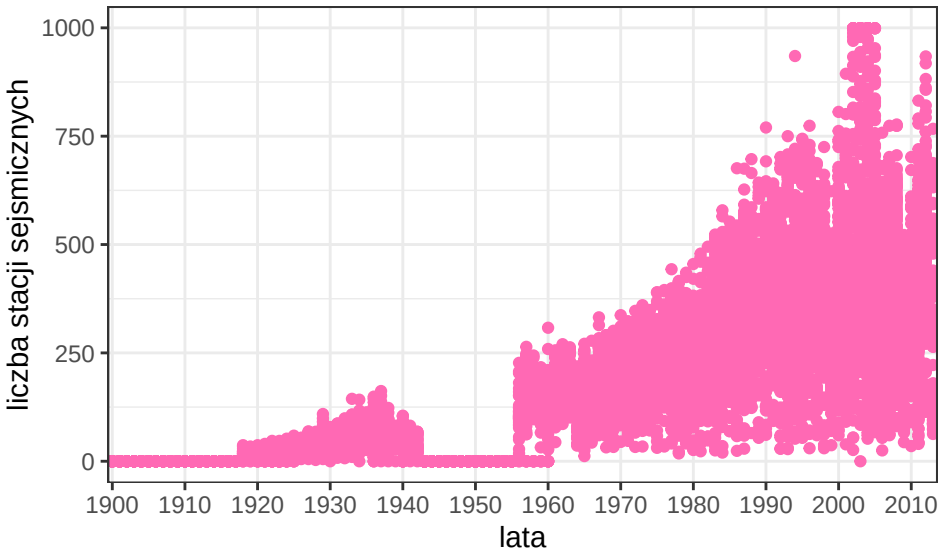
Wykres 2 potwierdza nasze wcześniejsze przypuszczenia oparte o wykres 1.

2.2 Macierz korelacji



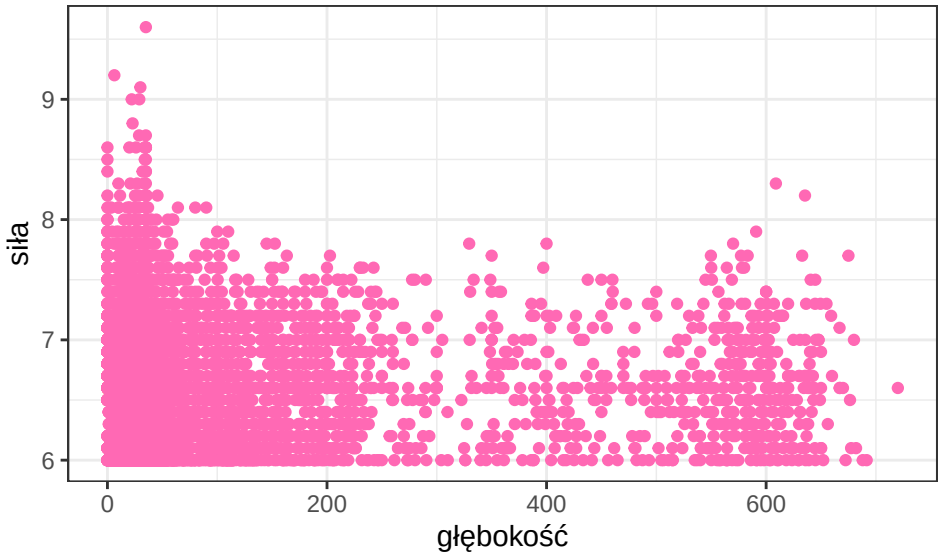
Wykres 3: Macierz korelacji

Jak widać na macierzy 3, najsilniejsza korelacja występuje pomiędzy kolumną *nst* oraz *years*. Dzięki temu możemy stwierdzić, że w ciągu kolejnych lat powstawało coraz więcej stacji sejsmicznych, co możemy pokazać bardziej szczegółowo, tworząc poniższy wykres:



Wykres 4: Scatterplot zależności liczby stacji sejsmicznych od lat.

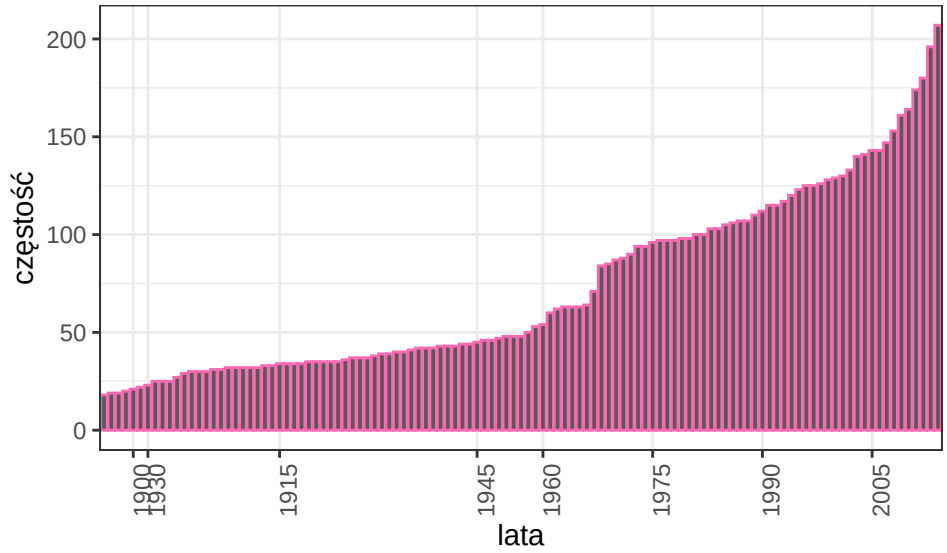
Zaskakiwać może fakt, że głębokość i siła trzęsienia nie są ze sobą silnie skorelowane, co możemy zobaczyć na wykresie poniżej:



Wykres 5: Scatterplot zależności siły trzęsienia od jego głębokości

Reszta kolumn nie jest ze sobą tak silnie skorelowana i nie jest tak zaskakująca jak zależności wyżej. Większość wartości oscyluje wokół zera, więc nie będziemy im się bliżej przyglądać.

Stosunek ilości trzęsień ziemi do kolejnych lat

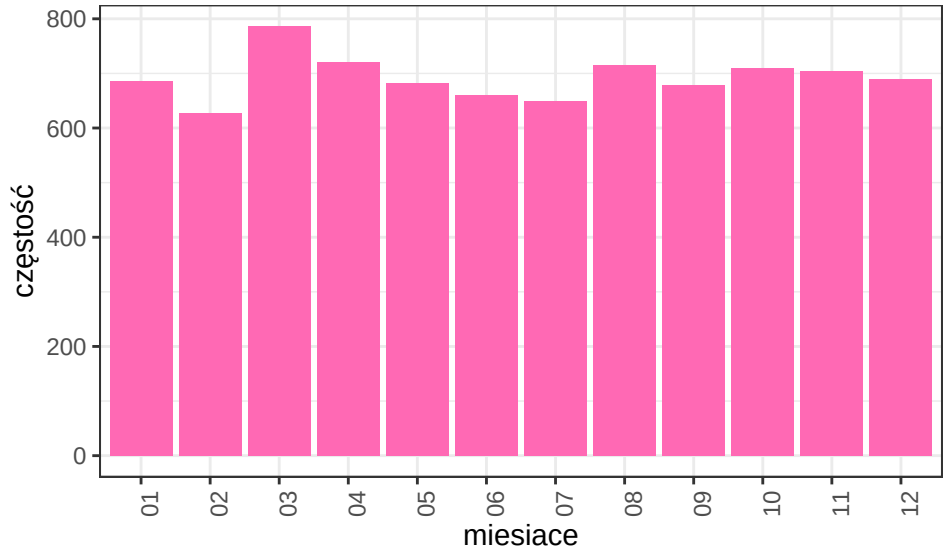


Wykres 6: Wykres kolumnowy ilości trzęsień ziemi w stosunku do lat

Na podstawie wykresu 6 można by przypuszczać, że wraz z upływem lat, przybywało trzęsień ziemi, jednak musimy wziąć pod uwagę fakt, że w tym samym czasie przybywało stacji sejsmicznych, które takie zdarzenia rejestrowały, co pokazaliśmy na wykresie 4. Zatem nie możemy jednoznacznie stwierdzić, czy liczba trzęsień ziemi zmieniała się w konretny sposób.

??? Czy to nie jest masło maślane XDDDDDDD Btw długie zdanie napisałam, a jak na polski się miało wymyślać to pustka XD co ta matma z człowiekiem robi. (sidenote: R dalej głupi)

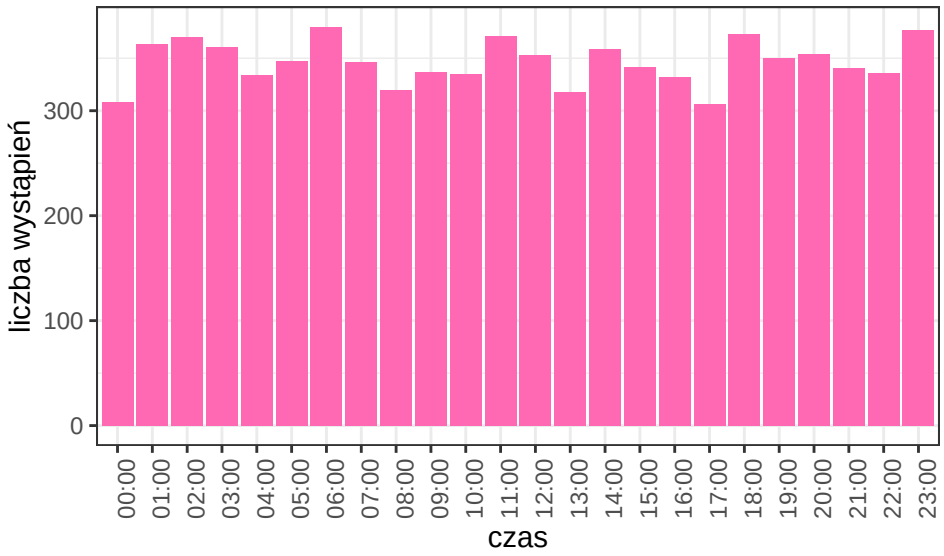
Powiedzmy że będę nazywać je dosadnie, jakkolwiek głupio to brzmi, a więc: ### Badanie wpływu pory roku na liczbę trzęsień ziemi



Wykres 7: Wykres kolumnowy ilości trzęsień ziemi w stosunku do miesiąca, w którym wystąpiły

Na powyższym wykresie możemy zauważyć, że słupki dla wszystkich miesięcy są dość podobnej wysokości. Nie ma żadnych mocno odstających liczb, zatem wnioskujemy, że pora roku nie ma wpływu na to, czy trzęsienie ziemi wystąpi, czy nie.

Badanie Zależności między porą dnia, a liczbą trzęsień ziemi



Wykres 8: Wykres kolumnowy ilości trzęsień ziemi w stosunku do godziny, w której wystąpiły? XD

Na potrzeby zrobienia wykresu, zaokrąglone, do najbliższej całkowitej godziny, zostały zmienne z kolumny *Time*. Tak jak przy badaniu wpływu pory roku, możemy zauważyć, że wartości znów są do siebie zbliżone, a zatem pora dnia również nie ma wpływu na występowanie trzęsień ziemi.

Nwm czy dodawać te boxploty bo one nie badają niczego w sumie, tylko obrazują zakres naszych danych więc można je wrzucic może do przedstawienia danych? XDDD Bez sensu trochę ale no, to w sumie wszystko co się nadaje wrzuciłam to by można jeszcze pomyśleć czy to już czy jak, bo tak nwm czy nie mało trochę XD

3. Wnioski