Katedra Elektrotechniki i Podstaw Informatyki

****

**Projekt ze statystycznej analizy danych – ludzie bezrobotni uprzednio pracujący w latach 2014 i 2024**

**Stechnij Damian**

Opiekun pracy:  
dr inż. Mariusz Startek

Rzeszów, 2024

Spis treści

[1. Wprowadzenie 3](#_Toc167659371)

[2. Wczytanie danych 4](#_Toc167659372)

[3. Wyznaczenie parametrów 6](#_Toc167659373)

[3.1. Średnia 6](#_Toc167659374)

[3.2. Wariancja 6](#_Toc167659375)

[3.3. Odchylenie standardowe 6](#_Toc167659376)

[3.4. Współczynnik zmienności 7](#_Toc167659377)

[3.5. Kwartyle, mediana, minimum, maksimum 7](#_Toc167659378)

[3.6. Dominanta 8](#_Toc167659379)

[3.7. Rozstęp 8](#_Toc167659380)

[3.8. Skośność 9](#_Toc167659381)

[3.9. Kurtoza 9](#_Toc167659382)

[4. Graficzna prezentacja danych 10](#_Toc167659383)

[4.1. Wykres pudełkowy 10](#_Toc167659384)

[4.2. Histogram 11](#_Toc167659385)

[4.3. Wykres dystrybuanty 12](#_Toc167659386)

[5. Hipotezy statystyczne 13](#_Toc167659387)

[5.1. Hipoteza dotycząca mediany liczby bezrobotnych 13](#_Toc167659388)

[5.2. Hipoteza dotycząca wariancji liczby bezrobotnych 13](#_Toc167659389)

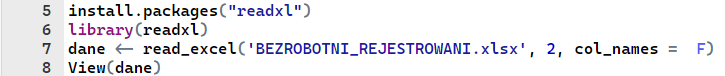
[6. Użyte biblioteki i polecenia 14](#_Toc167659390)

[7. Wnioski 15](#_Toc167659391)

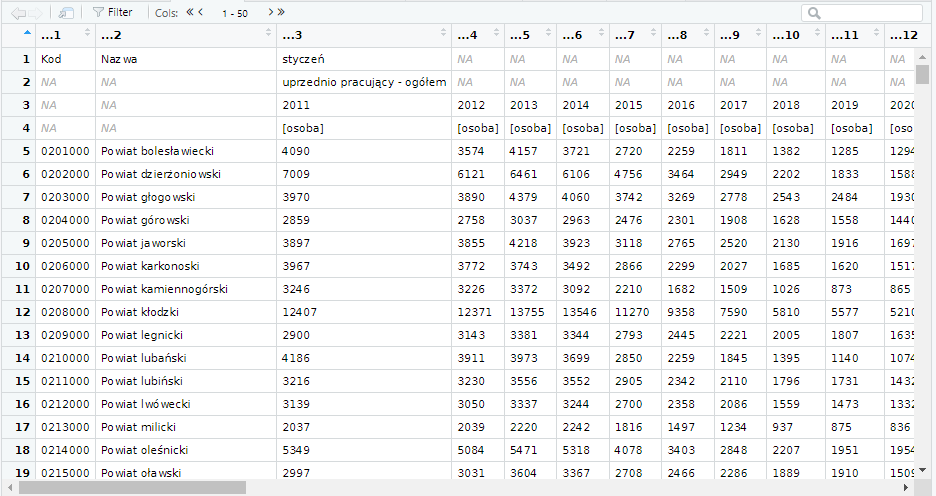
# Wprowadzenie

Projekt dotyczy statystycznej analizy danych ludzi bezrobotnych, którzy uprzednio pracowali. Dane zostały pobrane ze strony <https://bdl.stat.gov.pl>. Zawierają one lata od 2011 do 2024 z podziałem na miesiące oraz próbki ilości takowych ludzi pochodzą z powiatów w Polsce. Wybrano te dane, aby przeprowadzić analizę różnicy w próbkach pomiędzy rokiem 2014 a 2024 w miesiącu styczeń. W projekcie wyznaczono parametry opisowe, przedstawiono graficznie zebrane dane oraz zbadano hipotezy statystyczne.

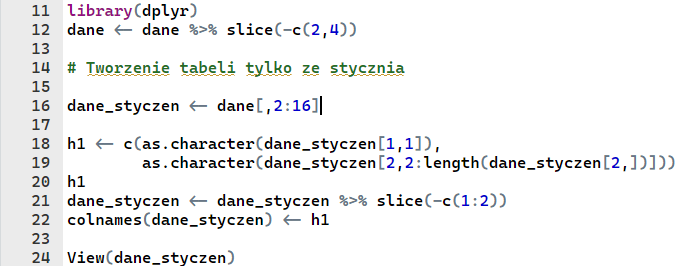
# Wczytanie danych

Wczytywany plik jest w formacie xlsx, więc do zaimportowania użyto biblioteki readxl i skorzystano z funkcji read\_excel.

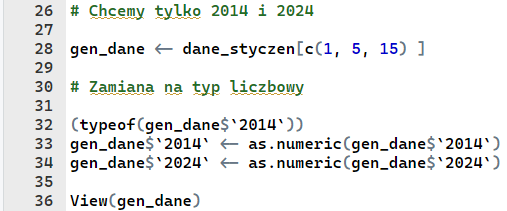
Zrzut ekranu 2.1 Wczytanie danych

Po wczytaniu danych otrzymujemy taką tabele z ilością osób bezrobotnych wcześniej pracujących z poszczególnych powiatów.

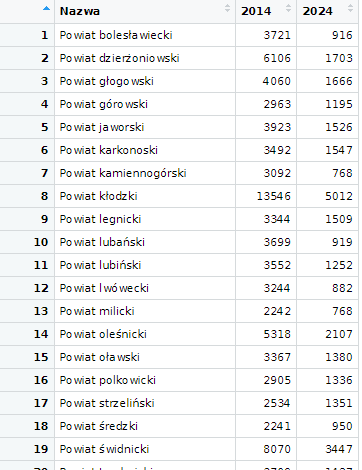
Zrzut ekranu 2.2 Tabela danych po wczytaniu

Postanowiono zebrać dane tylko ze stycznia, więc obcięto dane z pozostałych miesięcy oraz z pola z wartościami NA.

Zrzut ekranu 2.3 Obcięcie danych tylko do stycznia

Po obcięciu danych wybrano lata 2014 i 2024 by porównać 10 lat odstępu i zamieniono na typ liczbowy.

Zrzut ekranu 2.4 Wybranie tylko lat 2014 i 2024

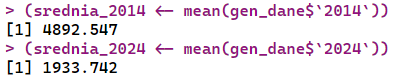
Gotowe dane po obróbce wyglądają następująco:

Zrzut ekranu 2.5 Dane po obróbce

# Wyznaczenie parametrów

## Średnia

Obliczono średnią arytmetyczną z obu lat. Służy ona do określania wartości centralnej zbioru danych.

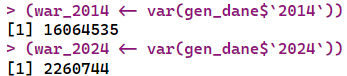


Zrzut ekranu 3.1 Obliczenie średniej

Średnia na przestrzeni lat zmieniła się ponad dwukrotnie, co oznacza, że o wiele mniej ludzi traci swoją pracę lub rzadziej zmieniają.

## Wariancja

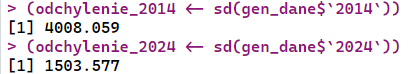
Następnie obliczono wariancję, która jest miarą rozproszenia danych wokół średniej wartości, dla obu kolumn. Wyniki mogą wysokie z powodu dużego rozrzutu wartości w danych.



Zrzut ekranu 3.2 Obliczenie wariancji

## Odchylenie standardowe

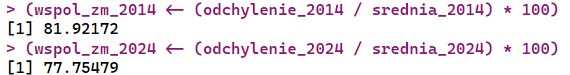
Odchylenie standardowe określa, ile wartość danej cechy odchyla się od obliczonej średniej arytmetycznej. W tym przypadku jej wartość jest bardzo zbliżone do samej średniej. Może to oznaczać, że wartości mogą wahać się nawet 2-krotnie od średniej.



Zrzut ekranu 3.3 Obliczenie odchylenia standardowego

## Współczynnik zmienności

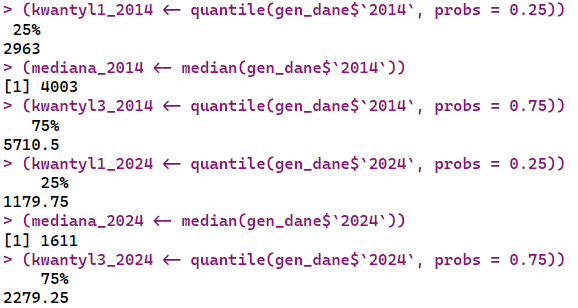
Poniżej są wyniki dla współczynnika zmienności, czyli dyspersji, dla obu lat. Dla 2014 wynosi blisko 82% a dla 2024 - 77,75%, co można zinterpretować jako znaczną dyspersję.



Zrzut ekranu 3.4 Obliczenie współczynnika zmienności

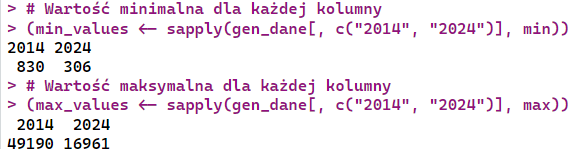
## Kwartyle, mediana, minimum, maksimum

Kwartyle dzielą zbiory na ćwiartki. Przykładowo pierwszy kwartyl oznacza, że 25% danych jest mniejszych niż Q1, a pozostałych 75% jest większych niż Q1. Mediana jest drugim kwartylem i dzieli badany zbiór na połowy.



Zrzut ekranu 3.5 Obliczenie kwantyli

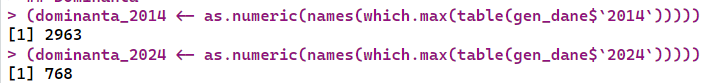
Poniżej znaleziono również wartości maksymalne jak i minimalne:



Zrzut ekranu 3.6 Znalezienie wartości maksymalnej i minimalnej

## Dominanta

Dominanta to wartość, którą wyznacza się poprzez znalezienie najczęściej występującej liczby w zbiorze danych.

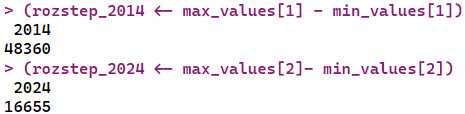


Zrzut ekranu 3.9 Obliczenie dominanty

Przedstawione wartości są całkiem niskie, więc może to oznaczać, że w wielu powiatach mało ludzi traci pracę, a na przestrzeni ostatnich 10 lat zmieniło się to niemal 4 krotnie.

## Rozstęp

Rozstęp to miara odległości jaką dzieli pomiędzy wartością maksymalną a minimalną obranej cechy statystycznej.

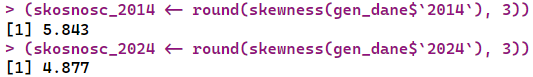


Zrzut ekranu 3.7 Obliczenie rozstępu

W roku 2014 jest znacząco wyższy niż w 2024, co może oznaczać, że w o wielu mniej miejscach ludzie są bezrobotni i jest ich mniej.

## Skośność

Skośność to miara asymetrii rozkładu zmiennej. Bada jak rozkład jest zbliżony do rozkładu normalnego.

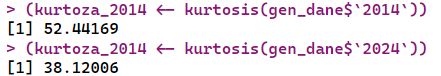


Zrzut ekranu 3.8 Obliczenie skośności

Dodatnie wyniki informują, że prawie ramię rozkładu jest wydłużone i wyniki poniżej średniej są przeważające w badanej próbce. Większa wartość w 2014 mówi o tym, że ramię jest bardziej wysunięte, niż 10 lat później.

## Kurtoza

Kurtoza jest miarą występowania wartości odstających. Podobnie jak skośność wskazuje jak bardzo rozkład jest zbliżony do rozkładu normalnego. Im bliżej wartości 0, tym bardziej analizowany rozkład jest zbliżony do normalnego.



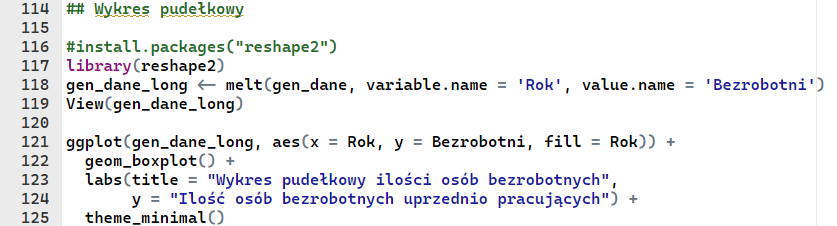
Zrzut ekranu 3.10 Wyniki dla kurtozy

Obliczona kurtoza jest większa od zera co wskazuje na letokurtyczność oraz znacząco odbiega co sugeruje, że dane zasadniczo różnią się od rozkładu normalnego. Tak wysokie wyniki wskazują na bardzo spiczasty szczyt oraz dłuższe i cieńsze ogony w porównaniu do rozkładu normalnego. Oznacza to, że jest większe prawdopodobieństwo wystąpienia wartości odstających.

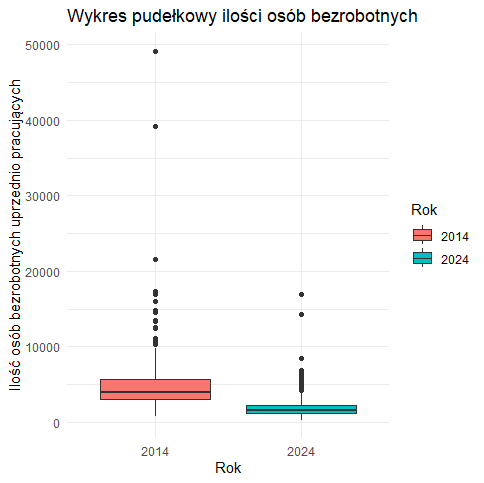
# Graficzna prezentacja danych

## Wykres pudełkowy

Na wykresie pudełkowy można przedstawić wiele informacji takie jak mediana, kwartyle, wartość minimum i maximum , przez co jest często używany.



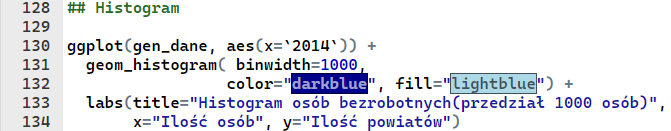
Zrzut ekranu 4.1 Kod tworzenia wykresu pudełkowego



Rysunek 4‑1 Wykres pudełkowy

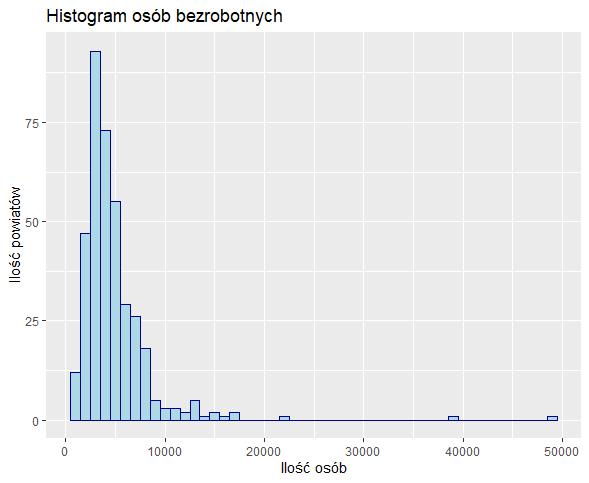
## Histogram

Histogram to wykres przedstawiający rozkładu empirycznego cechy. Jest podobny do wykresu słupkowego.



Zrzut ekranu 4.2 Kod tworzenia histogramu

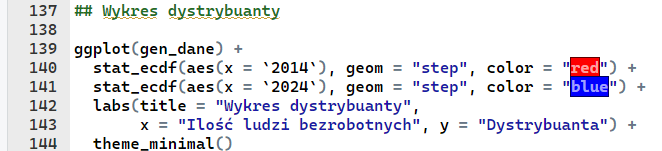
W kodzie ustawiono przedział, czyli szerokość słupka, na 1000. Poniżej znajduje się histogram.



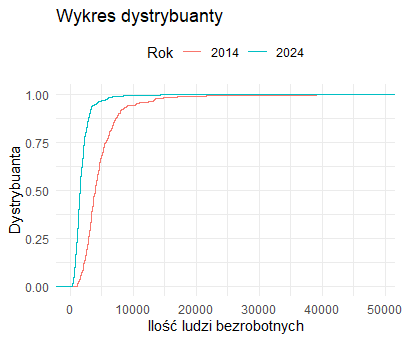
Rysunek 4‑2 Histogram

## Wykres dystrybuanty

Wykres dystrybuanty przedstawia rozkład prawdopodobieństwa występowania określonej wartości.



Zrzut ekranu 4.3 Kod tworzenia wykresu dystrybuanty



Rysunek 4‑3 Wykres dystrybuanty

# Hipotezy statystyczne

## Hipoteza dotycząca mediany liczby bezrobotnych

H0: Średnia ludzi bezrobotnych nie zmieniła się

H1: Średnia ludzi bezrobotnych zmniejszyła się

Wartość p jest mniejsza niż 0,05 co oznacza, że hipoteza zerowa jest nieprawdziwa

## Hipoteza dotycząca wariancji liczby bezrobotnych

# Użyte biblioteki i polecenia

# Wnioski