**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)

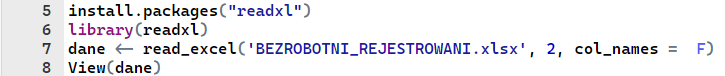
[6. Użyte biblioteki i polecenia 14](#_Toc167659390)

[7. Wnioski 15](#_Toc167659391)

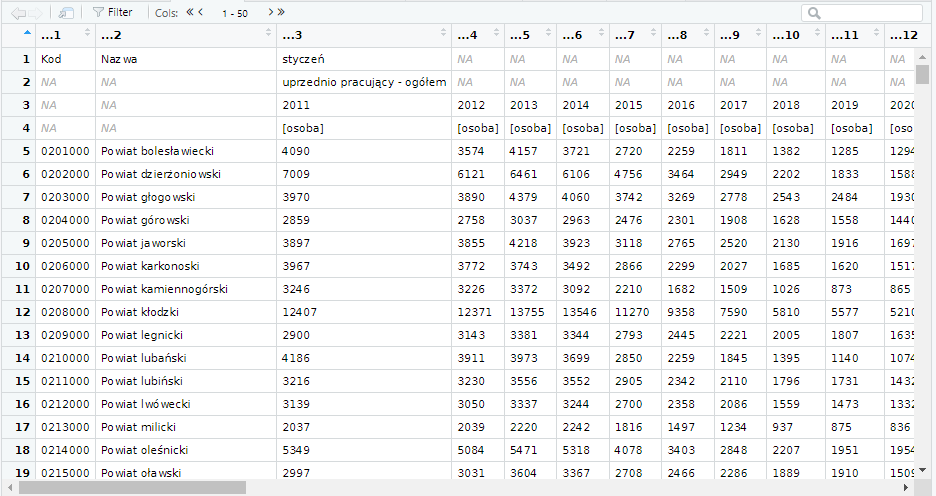
# Wprowadzenie

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.

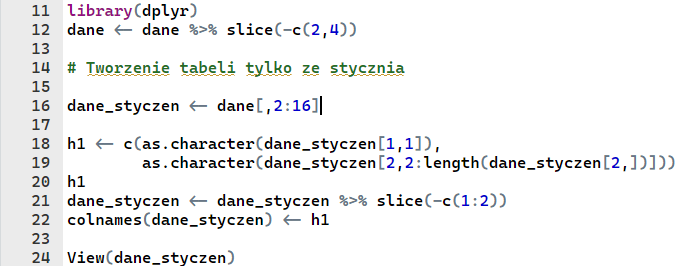
# Wczytanie danych

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

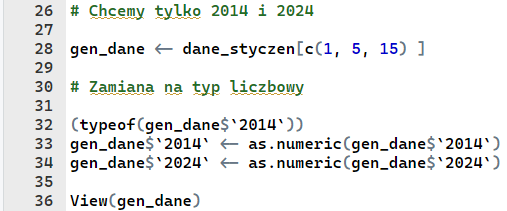
Zrzut ekranu 2.1 Wczytanie danych

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

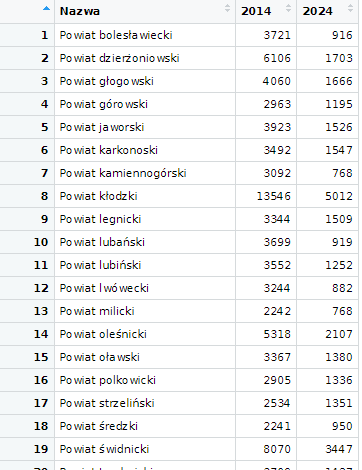
Zrzut ekranu 2.2 Tabela danych po wczytaniu

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

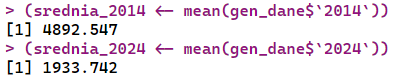
Gotowe dane po obróbce wyglądają następująco:

Zrzut ekranu 2.5 Dane po obróbce

# Wyznaczenie parametrów

## Średnia

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

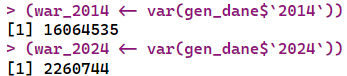


Zrzut ekranu 3.1 Obliczenie średniej

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

## Wariancja

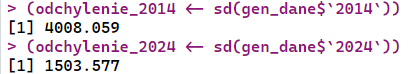
Następnie obliczono wariancję, która jest miarą rozproszenia danych wokół średniej wartości, dla obu kolumn. Wyniki mogą wysokie z powodu dużego rozrzutu wartości w danych.



Zrzut ekranu 3.2 Obliczenie wariancji

## Odchylenie standardowe

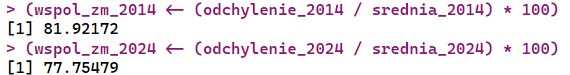
Odchylenie standardowe określa, ile wartość danej cechy odchyla się od obliczonej średniej arytmetycznej. W tym przypadku jej wartość jest bardzo zbliżone do samej średniej. Może to oznaczać, że wartości mogą wahać się nawet 2-krotnie od średniej.



Zrzut ekranu 3.3 Obliczenie odchylenia standardowego

## Współczynnik zmienności

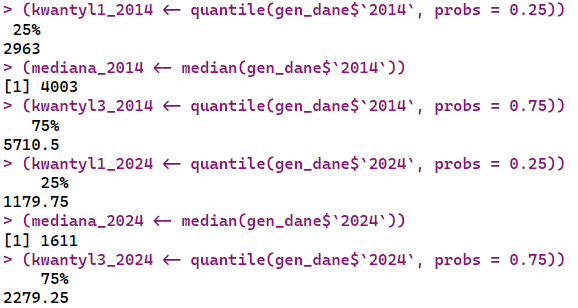
Poniżej są wyniki dla współczynnika zmienności, czyli dyspersji, dla obu lat. Dla 2014 wynosi blisko 82% a dla 2024 - 77,75%, co można zinterpretować jako znaczną dyspersję.



Zrzut ekranu 3.4 Obliczenie współczynnika zmienności

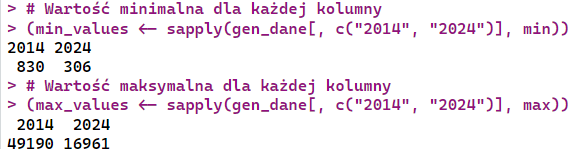
## Kwartyle, mediana, minimum, maksimum

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

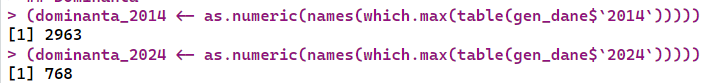
Poniżej znaleziono również wartości maksymalne jak i minimalne:



Zrzut ekranu 3.6 Znalezienie wartości maksymalnej i minimalnej

## Dominanta

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

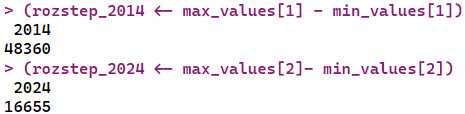


Zrzut ekranu 3.7 Obliczenie dominanty

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.

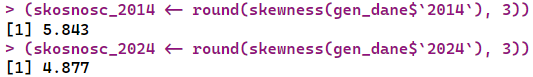


Zrzut ekranu 3.8 Obliczenie rozstępu

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. By móc użyć polecenia **skewness** użyto biblioteki „e1071”.

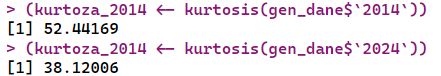


Zrzut ekranu 3.9 Obliczenie skośności

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.

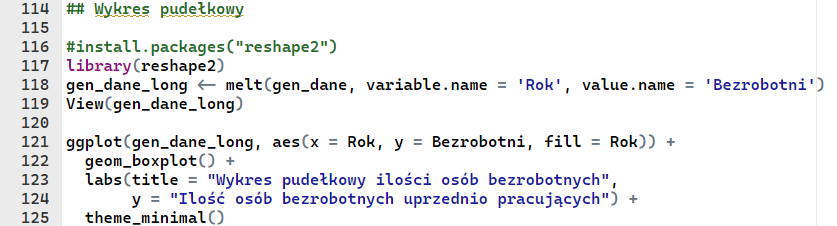
## Zestawienie wyników

|  |  |  |
| --- | --- | --- |
| **Rok** | **2014** | **2024** |
| **Średnia** | 4892,547 | 1933,742 |
| **Wariancja** | 16064535 | 2260744 |
| **Odchylenie standardowe** | 4008,059 | 1503,577 |
| **Współczynnik zmienności** | 81,922 % | 77,755 % |
| **Kwartyl 1. rzedu** | 2963 | 1179,75 |
| **mediana** | 4003 | 1611 |
| **Kwartyl 3. rzędu** | 5710,5 | 2279,25 |
| **Minimum** | 830 | 306 |
| **Maksymum** | 49190 | 16961 |
| **Dominanta** | 2963 | 768 |
| **Rozstęp** | 48360 | 16655 |
| **Skośność** | 5,843 | 4,877 |
| **Kurtoza** | 52,44169 | 38,12006 |

# Graficzna prezentacja danych

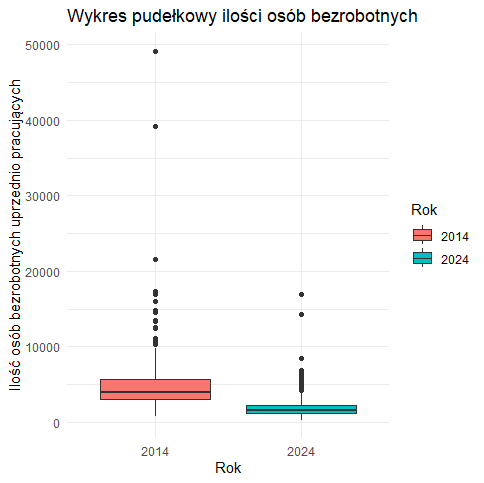
## Wykres pudełkowy

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

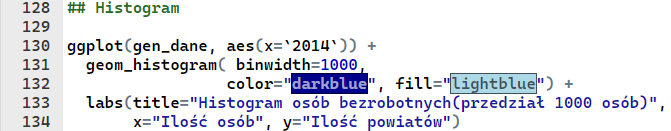
Na wykresach są bardzo duże wartości odstające. Tam, gdzie jest dolna część pudełka to jest to pierwszy kwartyl, a górna to trzeci kwartyl. Pozioma linia w pudełku to mediana. Wąsy to linie poza pudełkiem i ich końce mogą oznaczać od dołu wartość minimalna i od góry wartość maksymalną. W tym wypadku w górnej części wykresu są wartości odstające.



Rysunek 4‑1 Wykres pudełkowy

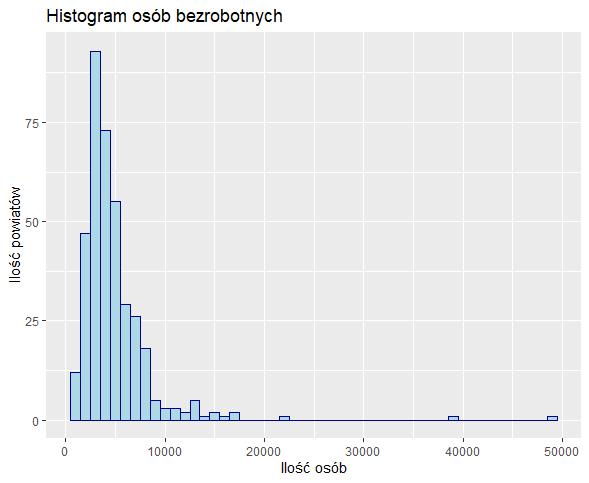
## Histogram

Histogram to wykres przedstawiający rozkładu empirycznego cechy. Jest podobny do wykresu słupkowego. Wysokość słupków, to ile elementów znajduje się w przedziale a szerokość słupków to przedział, w którym mieszczą się elementy.



Zrzut ekranu 4.2 Kod tworzenia histogramu

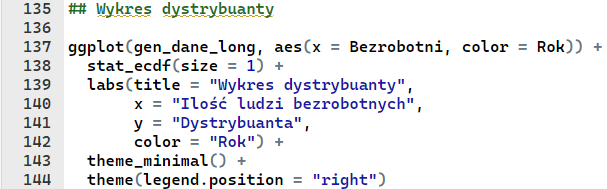
W kodzie ustawiono przedział, czyli szerokość słupka, na 1000. Poniżej znajduje się histogram.



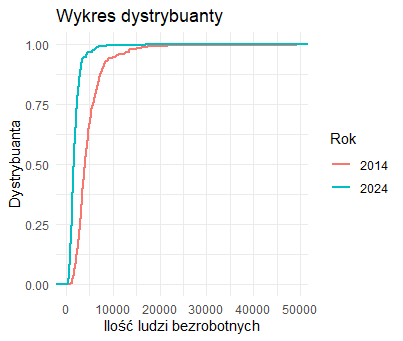
Rysunek 4‑2 Histogram

## Wykres dystrybuanty

Wykres dystrybuanty przedstawia rozkład prawdopodobieństwa występowania określonych wartości.



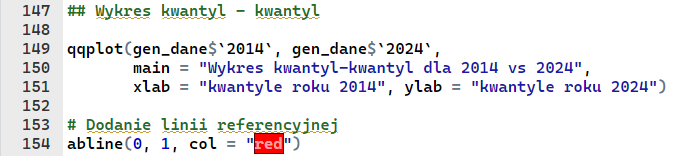
Zrzut ekranu 4.3 Kod tworzenia wykresu dystrybuanty



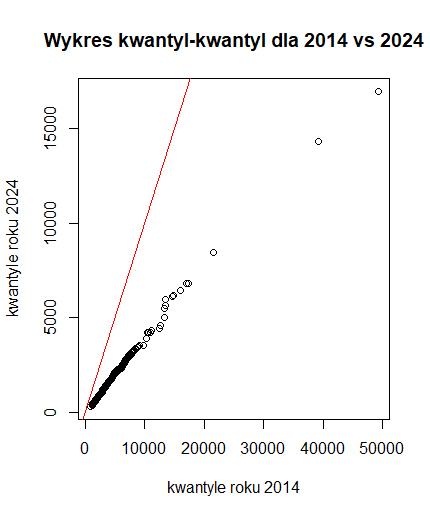
Rysunek 4‑3 Wykres dystrybuanty

## Wykres kwantyl-kwantyl

Wykres kwantyl-kwantyl można użyć do porównania danych z jakimś teoretycznym rozkładem. Służy do przedstawienia porównania kwantyli dwóch prób danych. W tym przypadku użyto do porównania obu prób danych i wizualizacji podobieństwa ich rozkładów co pozwala na ich ocenę.



Zrzut ekranu 4.4 Kod tworzenia wykresu kwantyl-kwantyl



Rysunek 4‑4 Wykres kwantyl-kwantyl

Czerwona linia to linia referencyjna o równaniu . Służy jako odniesienie do porównania rozkładów, pokazuje, gdzie leżałyby punkty, jeśli oba rozkłady były identyczne.

Na powyższym wykresie punkty leżą poniżej tej linii, co wskazuje, że wartości z roku 2024 są z reguły mniejsze niż te z roku 2014 dla odpowiadających kwantyli.

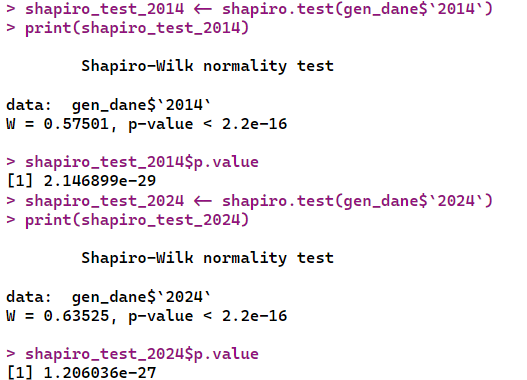
# Hipotezy statystyczne

## Hipoteza dotycząca rozkładu bezrobotnych

Drugą, poddaną testowi, zostaje hipoteza dotyczy tego czy dane posiadają rozkład normalny. Został do tego wykorzystany test Shapiro-Wilka.

**H0:** Dane pochodzą z rozkładu normalnego.

**H1:** Dane nie pochodzą z rozkładu normalnego.



Zrzut ekranu 5.1 Test Shapiro-Wilka

Wartość p wynosi 4.525356e-10 co znaczy ze jest bardzo bliskie 0 i jest znacznie mniejsze niż poziom istotności 0.05. Na tej podstawie odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną, która przyjmuje, że dane nie pochodzą z rozkładu normalnego.

## Hipoteza dotycząca mediany liczby bezrobotnych

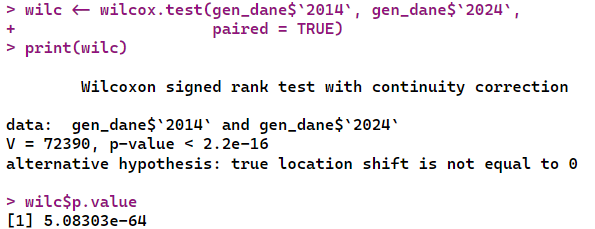
Do sprawdzenia hipotezy użyto nieparametrycznego testu Wilcoxona, który sprawdza, czy różnica pomiędzy medianami w dwóch próbach jest taka sama. Nasze dane są zależne, ponieważ zostały wykonane w tych samych powiatach, tyle że w odstępie czasowym 10 lat.

**H0:** Mediana ludzi bezrobotnych w powiatach nie różni się istotnie po 10 latach.

|  |  |
| --- | --- |
|  | Wzór 1 |

**H1:** Mediana ludzi bezrobotnych w powiatach różni się istotnie po 10 latach.

|  |  |
| --- | --- |
|  | Wzór 2 |



Zrzut ekranu 5.2 Test Wilcoxona

Wartość p jest bardzo bliska zeru i jest mniejsza niż poziom istotności 0,05. Na tej podstawie odrzucamy hipotezę zerową i możemy uznać hipotezę alternatywną za prawdziwą.

# Użyte biblioteki i polecenia

## Do wczytania danych

Do wczytania danych użyto biblioteki „readxl”, dzięki której łatwiej wyciągnąć dane z Excela do R. Przy pomocy polecenia **read\_excel()** można odcztać pliki o rozszerzeniu .xls, .xlsx.

Przy obróbce danych wykorzystano bibliotekę „dplyr”, by skorzystać z operatora **%>%**, by łatwiej było odcinać niepotrzebne kolumny i wiersze.

**As.numeric()** – zamienia dane na typ liczbowy

## Do obliczeń parametrów

**mean()** – oblicza średnią arytmetyczną

**var()** – oblicza wariancję

**sd()** – oblicza odchylenie standardowe

**quantile()** – oblicza kwantyle, dzięki **probs** określa, jak mają zostać podzielone dane

**mediane()** – znajduje medianę

**sapply(dane, min/max)** – znajduje wartości minimalne lub maksymalne

**skewness()** – oblicza skośność, pochodzi z biblioteki „e1071”

**kurtosis()** – oblicza kurtozę, również z biblioteki „e1071”

Dzięki **which.max()** znaleziono dominantę.

## Do wykresów

Użyto bibliotekę „reshape2”, by przy pomocy polecenia melt() móc przekształcić dane, dzięki czemu łatwiej można było zaimplementować je do wykresów.

Aby utworzyć wykresy wykorzystano bibliotekę „ggplot2”.

**ggplot(data = df, mapping = aes(x, y, other aesthetics)** – inicjuje obiekt wykres,

**aes() –** określa co jest jako x i co jako y

**geom\_boxplot()** – tworzy wykres pudełkowy

**geom\_histogram()** – nakłada warstwę z histogramem

**stat\_ecdf()** – tworzy wykres dystrybuanty

**labs()** – tworzy etykiety do legend

**theme\_minimal()** – ustawia minimalny styl wykresu

## Do hipotez

**wilcox.test()** – wykonuje test Wilcoxona, parametr paired = TRUE dla danych sparowanych

**shapiro.test()** – wykonuje test Shapiro-Wilka, by zbadać, czy dane mają rozkład normalny

# Wnioski

Analizując dane o bezrobotnych, którzy uprzednio pracowali po odstępie 10 lat, pomiędzy rokiem 2014 a 2024 w miesiącu styczeń, można wyciągnąć pewne wnioski.

Rynek pracy się zmienił na tyle, by o wiele mniej ludzi traciło pracę lub ludzie o wiele chętniej zostają w swojej pracy. Po wszelkich parametrach było widać, że te dane są znacząco mniejsze, bo nawet ponad dwukrotnie. Średnie po takim odstępie czasu znacząco się różniły.

Po graficznym przedstawieniu danych poprzez histogram czy wykres pudełkowy można zaobserwować asymetrie lewostronną, która na przestrzeni lat jeszcze bardziej się zmieniła. Można również stwierdzić, że rozrzut danych się istotnie zmniejszył.

Analiza ta pomaga w dostrzeżeniu zmian na polskim rynku pracy.