AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

Wydział Zarządzania

# Ekonometria

Projekt 4

Temat: Wstępna analiza danych służących do stworzenia modelu regresji wielorakiej

Autorzy: Laura Cygan, Wiktoria Bąk

Wydział: Wydział Zarządzania

Kierunek: Informatyka i Ekonometria

Przedmiot: Ekonometria

**Kraków, 2024**

# I. Przedstawienie problemu

Celem naszego projektu było wykonanie wstępnej analizy wybranych przez nas danych, które następnie posłużą do stworzenia modelu regresji wielorakiej. Jako zmienną objaśnianą do naszego modelu wybraliśmy bezrobocie w Polsce. W celu próby wyjaśnienia, co wpływa na procent niezatrudnionych ludzi w poszczególnych powiatach wybraliśmy pięć potencjalnych zmiennych objaśniających: średnie wynagrodzenie, współczynnik feminizacji, współczynnik urbanizacji, liczbę ofert pracy i liczbę małżeństw. Dane ujęliśmy dla wszystkich powiatów ogółem, jak i również z podziałem na te grodzkie (miasta na prawach powiatu) i ziemskie. Wszystkie dane pochodzą z Bazy Danych Lokalnych GUS i przedstawiają sytuację na 2021 rok.

# II. Opis zmiennych

**Bezrobocie** – przedstawione jako procent populacji danego powiatu. Z definicji, "Pojęcie osoby bezrobotnej oznacza ogólnie osobę niezatrudnioną, nieprowadzącą działalności gospodarczej i niewykonującą innej pracy zarobkowej, zdolną i gotową do podjęcia zatrudnienia."[1] Wybraliśmy tę zmienną, aby zbadać, jak wygląda sytuacja na rynku pracy w Polsce i jak różni się ona w zależności od powiatu.

**Wynagrodzenie** – Przeciętne wynagrodzenie w danym powiecie podane w złotówkach. Ta miara odzwierciedla poziom wynagrodzeń pracowników w danym obszarze i jest jednym z kluczowych wskaźników ekonomicznych, które mogą wpływać na bezrobocie.

**Współczynnik feminizacji** - Określa, ile kobiet w danym powiecie przypada na 100 mężczyzn. Przedstawiony jest w procentach. Wysoki współczynnik feminizacji może wskazywać na dominację kobiet w danej populacji, co ma potencjalny wpływ sytuację ekonomiczną powiatu.

**Współczynnik urbanizacji** - procentowy udział mieszkańców miast w ogólnej liczbie ludności powiatu. Współczynnik ten jest wskaźnikiem rozwoju społeczeństwa. Jego wysoka wartość może świadczyć o stopniu rozwoju infrastruktury miejskiej, dostępności usług oraz zatrudnienia.

[1] *Bezrobocie,* Wikipedia, https://pl.wikipedia.org/wiki/Bezrobocie [dostęp 26.04.2024 r.]

**Oferty pracy** – ilość ofert pracy zaoferowanych w ciągu roku, przypadających na 10 tys. mieszkańców danego powiatu. Ta zmienna może być istotnym wskaźnikiem aktywności ekonomicznej danego regionu oraz zapotrzebowania na siłę roboczą.

**Małżeństwa** – liczba zawartych małżeństw w danym powiecie w ciągu roku, przypadająca na 10 tys. mieszkańców. Zawieranie związków małżeńskich przez mieszkańców powiatu może mieć potencjalny wpływ na ich chęć oraz możliwość podjęcia pracy.

# DLA WSZYSTKICH POWIATÓW

## III – V. Wstępna analiza danych

### STATYSTYKI OPISOWE

Naszą analizę rozpoczęłyśmy od zbadania podstawowych statystyk opisowych wybranych przez nas danych.

BEZROBOCIE:

| Średnia | Odchylenie st. | Skośność | Kurtoza |
|---|---|---|---|
| 2.749847 | 1.328135 | 1.028822 | 0.9548021 |

Średnie bezrobocie we wszystkich powiatach wynosi ok. 2,75%. Oznacza to, że na każde 100 osób ok. 2-3 osoby są bezrobotne.

Odchylenie standardowe wynosi ok. 1,33. Oznacza to, że od średniej wartości bezrobocia (2,75%) wartości w poszczególnych powiatach mogą różnić się o ok. 1,33 punkty procentowe.

Skośność na poziomie 1,03 wskazuje na prawostronną asymetrię rozkładu. Oznacza to, że więcej powiatów ma poziom bezrobocia poniżej obliczonej średniej.

Kurtoza powyżej zera (0,95) wskazuje na rozkład leptokurtyczny, co oznacza, że wartości cechy są skoncentrowane wokół średniej oraz istnieje większa szansa na pojawienie się outlierów.

| | |
|---|---|
| Współczynnik zmienności | 48.29849 |

Współczynnik zmienności to iloraz odchylenia standardowego cechy i jej średniej arytmetycznej. Wartość na poziomie ok. 48,3% wskazuje na zmienność przeciętną, w stronę silnej. Oznacza to całkiem duże zróżnicowanie cechy i prawdopodobną niejednorodność badanej populacji. Miara ta jest szczególnie ważna przy doborze zmiennych objaśniających do modelu.

WYNAGRODZENIA:

| Średnia | Odchylenie st. | Skośność | Kurtoza |
|---|---|---|---|
| 5213.844 | 640.554 | 2.663638 | 12.37844 |

Przeciętne wynagrodzenie w powiatach w Polsce wynosi około 5213,84 złotych.

Odchylenie standardowe wynosi ok. 640,55 złotych. Oznacza to, że od średniej wartości wynagrodzeń wartości wynagrodzeń w powiatach mogą różnić się o około 640,55 zł.

Skośność wynosi około 2,66. Dodatnia skośność oznacza, że rozkład wynagrodzeń jest skoncentrowany na lewo od średniej, co sugeruje, że jest więcej powiatów z niższymi zarobkami.

Kurtoza wynosi około 12,38. Jest to bardzo wysoka wartość dodatnia, która wskazuje na to, że rozkład wynagrodzeń jest skoncentrowany wokół średniej i może sugerować występowanie wartości odstających.

| | |
|---|---|
| Współczynnik zmienności | 12.28564 |

Współczynnik zmienności wynosi około 12.29%. Wskazuje to na stosunkowo niską zmienność danych. Zmienne objaśniające w modelu powinny posiadać wartość V > 10%, co w tym przypadku jest spełnione.

WSPÓŁCZYNNIK FEMINIZACJI:

| Średnia | Odchylenie st. | Skośność | Kurtoza |
|---|---|---|---|
| 105.0053 | 3.989372 | 1.047879 | 0.6648558 |

Średni współczynnik feminizacji w badanej populacji wynosi ok. 105%. Oznacza to, że na każde 100 mężczyzn przypada średnio 105 kobiet.

Odchylenie standardowe wynosi ok. 3.99. Oznacza to, że od średniej wartości współczynnika feminizacji (105%) wartości współczynnika feminizacji dla konkretnych powiatów mogą różnić się o około 3.99 punkty procentowe.

Skośność na poziomie 1,04 sugeruje prawostronną skośność rozkładu. Oznacza to, że większość powiatów posiada poziom współczynnika feminizacji poniżej wyznaczonej średniej (105%).

Dodatnia kurtoza na poziomie 0,66 wskazuje na leptokurtyczność. Rozkład współczynnika feminizacji ma ogony, które są "cięższe" w porównaniu do tych w rozkładzie normalnym. Oznacza to, że istnieje większe prawdopodobieństwo wystąpienia wartości odstających.

| | |
|---|---|
| Współczynnik zmienności | 3.799211 |

Wartość współczynnika zmienności na poziomie ok. 3,8% wskazuje na małą zmienność danych i jest zdecydowanie niezadawalająca w kontekście tworzenia modelu regresji (nie spełnia warunku V > 10%).

WSPÓŁCZYNNIK URBANIZACJI:

| Średnia | Odchylenie st. | Skośność | Kurtoza |
|---|---|---|---|
| 50.68571 | 27.26119 | 0.6323097 | 0.9548021 |

Średni współczynnik urbanizacji wynosi ok. 50,69%. Oznacza to, że około połowa populacji powiatów zamieszkuje tereny miejskie.

Odchylenie standardowe wynosi ok. 27,26. Oznacza to, że od średniej wartości współczynnika urbanizacji wartości współczynnika urbanizacji w powiatach mogą różnić się o około 27,26 p.p.

Skośność wynosi około 0,63. Prawostronna skośność sugeruje, że więcej obszarów ma niższy stopień urbanizacji niż średnia.

Kurtoza na poziomie 0,95 (K > 0) wskazuje na rozkład leptokurtyczny, co oznacza, że wartości współczynnika urbanizacji są skoncentrowane wokół średniej oraz istnieje większa szansa na pojawienie się wartości odstających.

| | |
|---|---|
| Współczynnik zmienności | 53.78475 |

Wartość współczynnika zmienności wynosi 53,78%. Wskazuje ona na silną zmienność i zróżnicowanie danych. Spełnia warunek doboru zmiennych objaśniających do modelu (V > 10%). Tak wysoki współczynnik zmienności może wynikać z uwzględnienia w danych zarówno powiatów ziemskich, jak i grodzkich (w których współczynnik urbanizacji wynosi 100%).

OFERTY PRACY:

| Średnia | Odchylenie st. | Skośność | Kurtoza |
|---|---|---|---|
| 343.4768 | 192.4349 | 1.614611 | 3.061895 |

Średnia liczba ofert pracy w ciągu roku wynosi około 343,48. Oznacza to, że przeciętnie na 10 tysięcy mieszkańców powiatu przypadają ok. 343 oferty pracy.

Odchylenie standardowe wynosi ok. 192,43. Oznacza to, że od średniej wartości liczby ofert pracy (na 10 tys. mieszkańców) wartości liczby ofert pracy w powiatach mogą różnić się o około 192 oferty.

Skośność wynosi ok. 1,61, co oznacza prawostronną skośność rozkładu. Sugeruje to, że więcej obszarów ma niższą liczbę ofert pracy niż średnia.

Kurtoza wynosi ok. 3,06. Oznacza to, że rozkład wynagrodzeń jest skoncentrowany wokół średniej i może posiadać wartości odstające.

| | |
|---|---|
| Współczynnik zmienności | 56.02559 |

Wartość współczynnika zmienności wynosi 56%. Wskazuje to na silną zmienność danych i oznacza, że liczba ofert pracy wstępnie nadaje się na zmienną objaśniającą modelu.

MAŁŻEŃSTWA:

| Średnia | Odchylenie st. | Skośność | Kurtoza |
|---|---|---|---|
| 43.25969 | 4.564734 | 0.1284776 | 0.2367646 |

Średnia liczba zawartych małżeństw wynosi ok. 43,26. Oznacza to, że w powiatach średnio na każde 10 tysięcy mieszkańców przypadają 43 zawarte małżeństwa w ciągu roku.

Odchylenie standardowe wynosi około 4,56. Oznacza to, że od średniej wartości liczby zawartych małżeństw (43,26) wartości liczby małżeństw w populacji mogą różnić się o około 4,56.

Skośność wynosi ok. 0,13. Jest ona bliska zeru, co sugeruje, że rozkład liczby zawartych małżeństw jest stosunkowo symetryczny. To oznacza, że liczba małżeństw w powiatach jest równomiernie rozłożona względem średniej wartości.

Kurtoza wynosi około 0,24. Wartość bliska zeru sugeruje, że rozkład liczby zawartych małżeństw jest zbliżony do rozkładu normalnego.

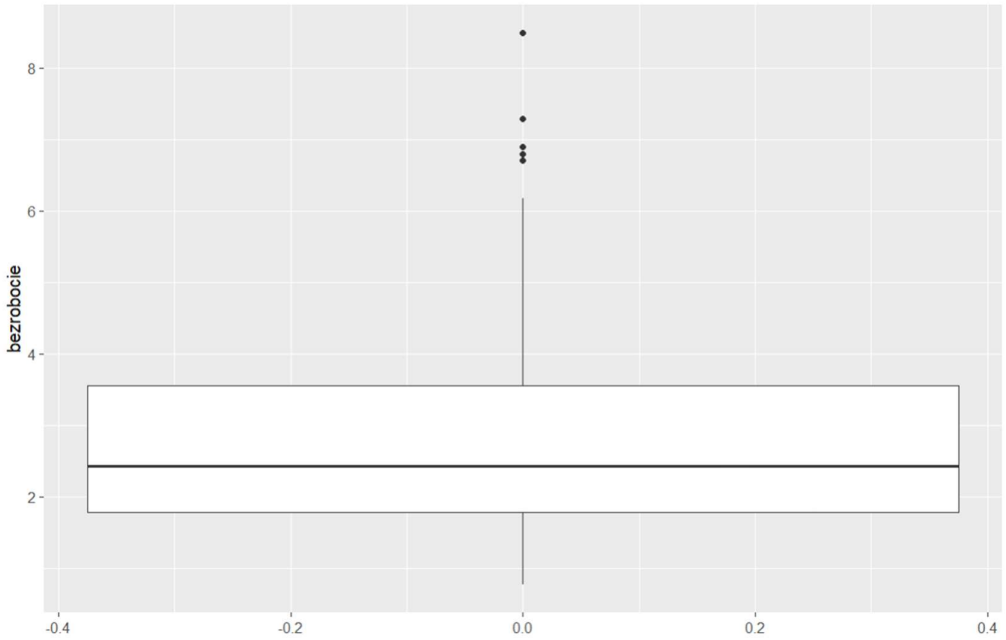
| Współczynnik zmienności | 10.55193 |
|---|---|

Współczynnik zmienności wynosi ok. 10,55%. Oznacza to raczej słabą zmienność danych, jednak jeszcze akceptowalną dla zmiennej objaśniającej modelu regresji (V > 10%).

## WYKRESY PUDEŁKOWE

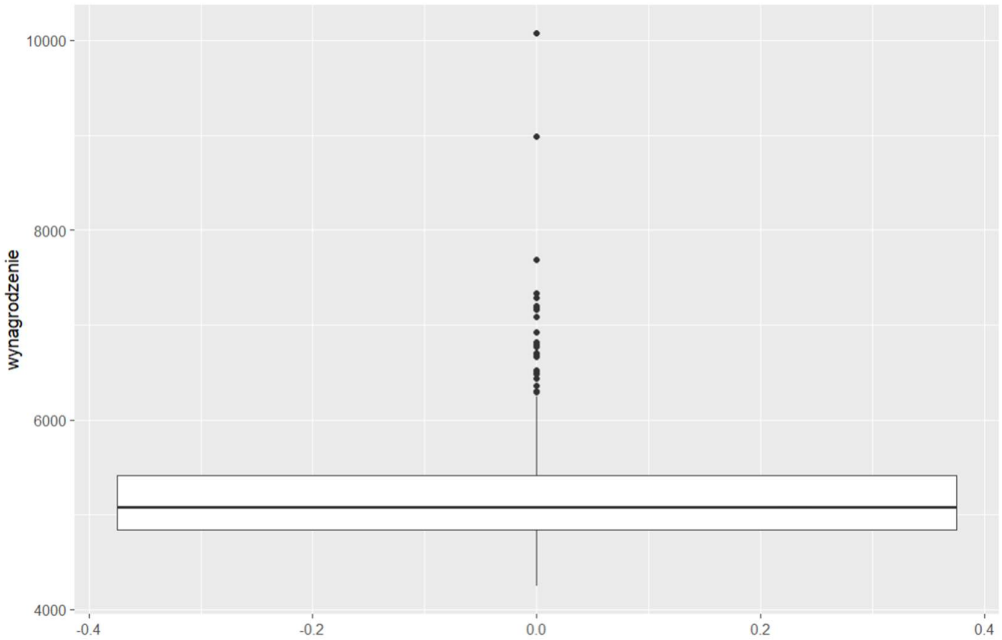
Następnie postanowiłyśmy narysować wykresy pudełkowe poszczególnych zmiennych i zobaczyć, jak wizualnie prezentują się ich rozkłady.

BEZROBOCIE



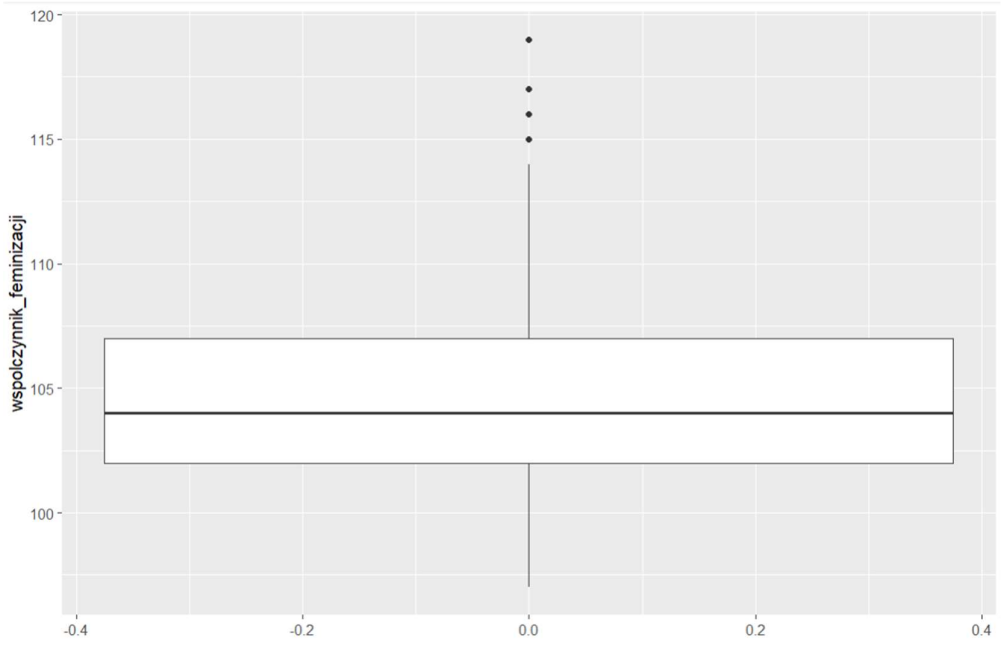

Wnioski jakie można wyciągnąć z wykresu rozkładu bezrobocia pokrywają się z tymi wyciągniętymi za pomocą statystyk opisowych. Rozkład jest prawostronnie skośny i posiada parę wartości odstających. Górny wąs wykresu jest zdecydowanie dłuższy niż dolny.

WYNAGRODZENIE



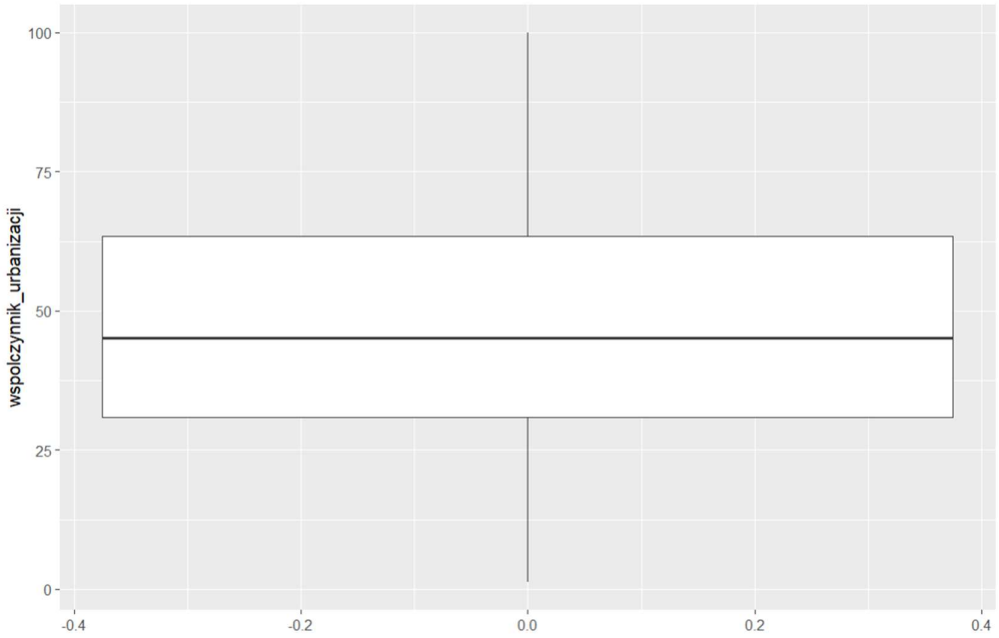
W tym przypadku również rozkład jest prawostronnie skośny. Dane dotyczące wynagrodzenia posiadają dużo wartości odstających, co prawdopodobnie wynika ze zdecydowanie wyższego poziomu średniego wynagrodzenia w dużych miastach.

WSPÓŁCZYNNIK FEMINIZACJI



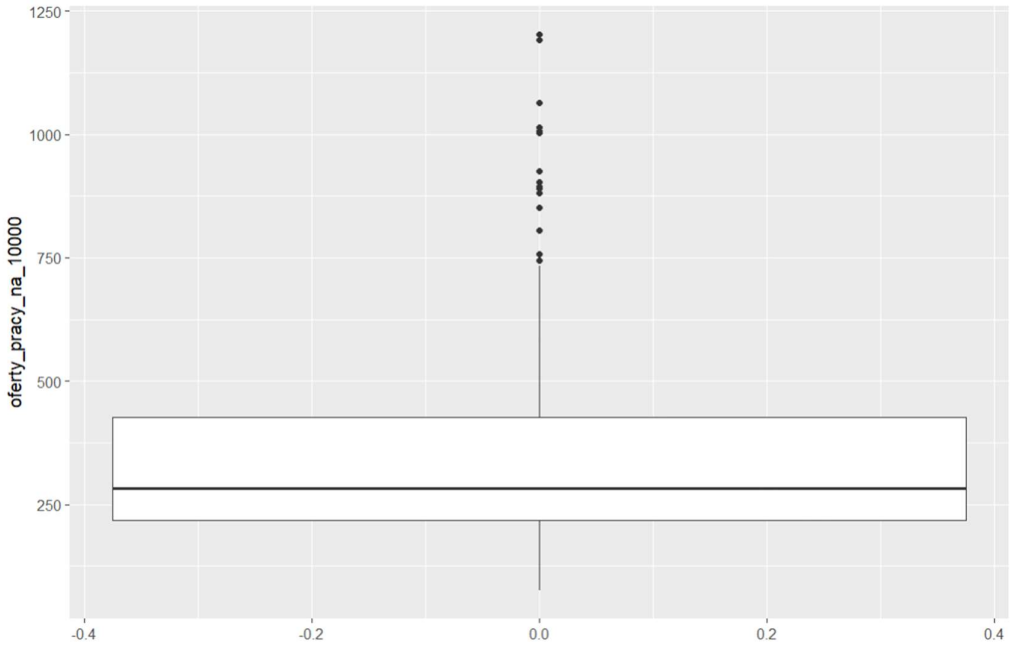

Współczynnik feminizacji również posiada wykres prawoskośny oraz parę wysokich wartości odstających,

WSPÓŁCZYNNIK URBANIZACJI



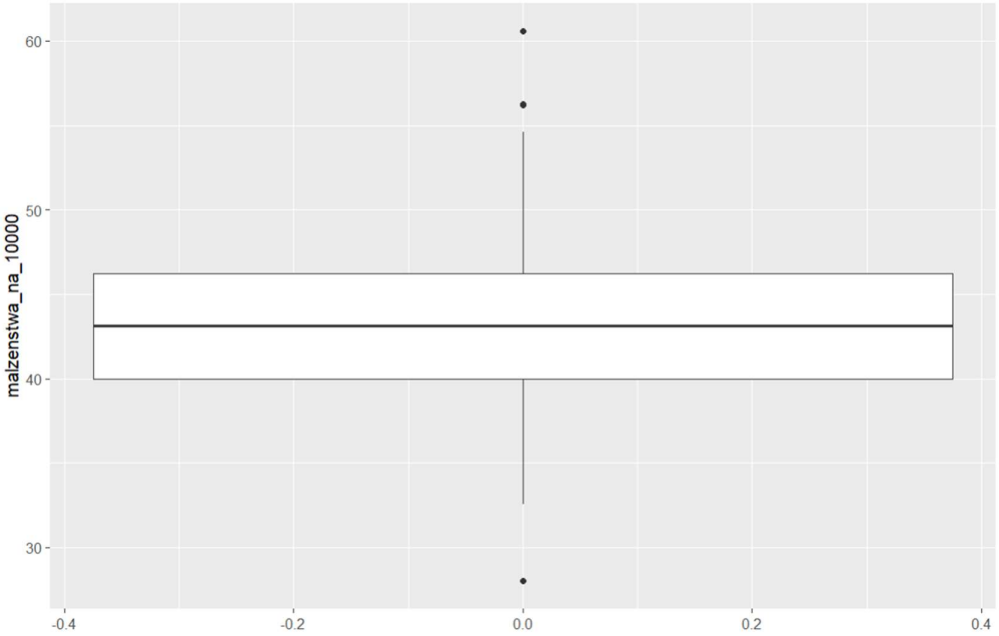

Rozkład współczynnika urbanizacji może przypominać rozkład normalny, jednak z statystyk opisowych wiemy, że nie jest to do końca prawda. Również z wykresu możemy zobaczyć, że mediana współczynnika urbanizacji plasuje się na poziomie poniżej 50.

OFERTY PRACY



Oferty pracy mają wyraźnie prawoskośny rozkład. Widocznych jest również dużo outlierów o wysokich wartościach, co ponownie, może wynikać z np. większej ilości ofert pracy w dużych miastach.

LICZBA MAŁŻEŃSTW




Liczba małżeństw ma rozkład podobny do rozkładu normalnego. Możemy zaobserwować niewiele wartości odstających.

Pomimo występowania wartości odstających w rozkładach (szczególnie przy przeciętnym wynagrodzeniu i liczbie ofert pracy) postanowiliśmy nie usuwać nietypowych obserwacji. Są one wynikiem rzeczywistych zdarzeń w powiatach, co ma znaczenie dla analizy. Usunięcie ich może prowadzić do selektywnego interpretowania danych oraz zniekształcenia obrazu rzeczywistości.

# VI. KORELACJE

W kolejnym etapie policzyłyśmy korelacje między wszystkimi zmiennymi (zarówno zmienną objaśnianą z objaśniającymi, jak i objaśniającymi między sobą). Jest to kolejny sposób na dobór odpowiednich zmiennych do modelu regresji.

| | Bezrobocie | Wynagrodzenie | Ws. feminizacji | Ws. urbanizacji | Oferty pracy | Małżeństwa |
|---|---|---|---|---|---|---|
| Bezrobocie | 1.0 | -0.2789 | -0.2615 | -0.2482 | -0.1923 | -0.0995 |
| Wynagrodzenie | -0.2789 | 1.0 | 0.4696 | 0.4487 | 0.2020 | 0.0831 |
| Ws. feminizacji | -0.2615 | 0.4696 | 1.0 | 0.8225 | 0.2443 | -0.1313 |
| Ws. urbanizacji | -0.2482 | 0.4487 | 0.8225 | 1.0 | 0.2542 | -0.1144 |
| Oferty pracy | -0.1923 | 0.2020 | 0.2443 | 0.2542 | 1.0 | -0.0584 |
| Małżeństwa | -0.0995 | 0.0831 | -0.1313 | -0.1144 | -0.0584 | 1.0 |

Z założenia, zmienne objaśniające powinny mieć silną korelację ze zmienną objaśnianą (niezależnie od znaku), a być nieskorelowane między sobą. Jak widać, żadna zmienna nie wykazuje mocnej korelacji z bezrobociem – najwyższe wartości osiągane przez wynagrodzenie (-0.28), współczynnik feminizacji (-0.26) i współczynnik urbanizacji (-0.25) świadczą o zaledwie słabej zależności. Sytuacja nie wygląda korzystnie również patrząc na korelacje między zmiennymi objaśniającymi – pomiędzy współczynnikiem feminizacji a współczynnikiem urbanizacji korelacja wynosi aż 0.82. Również między wynagrodzeniem a wskaźnikami feminizacji i urbanizacji możemy zaobserwować umiarkowaną zależność (odpowiednio 0.47 i 0.45). Najmniejsze skorelowanie z innymi zmiennymi wykazuje liczba małżeństw.

# VII. METODA HELLWIGA

Za pomocą metody Hellwiga postanowiłyśmy wyznaczyć najlepszy zestaw zmiennych objaśniających poziom bezrobocia. W tym celu wykorzystałyśmy kod napisany w RStudio.

```
Największa wartość integralnej pojemności informacyjnej: 0.1137819
Najlepszy podzbiór zmiennych: X1  X2  X4  X5
```

Program jako najlepszy zestaw zmiennych objaśniających wybrał przeciętne wynagrodzenie, współczynnik feminizacji, liczbę ofert pracy i liczbę zawartych małżeństw (oznaczonych jako X1, X2, X3 i X5). Integralna pojemność informacyjna dla tego zestawu wynosi ok. 0.11, co jest najwyższą dostępną, jednak wciąż względnie niską wartością.

### WNIOSKI

Dla wszystkich zmiennych zaobserwowaliśmy rozkłady prawostronnie skośne, z kurtozą wyższą od rozkładu normalnego. Mimo występowania wartości odstających zdecydowaliśmy się ich nie usuwać, aby nie utracić istotnych informacji i zachować wiarygodność modelu. Za niską wartość współczynnika zmienności w kontekście tworzenia modelu regresji posiadał jednie współczynnik feminizacji (3,8%). Korelacje zmiennej objaśnianej (bezrobocia) ze zmiennymi objaśnianymi są dosyć niskie – najwyższa wartość wynosiła ok. -0.28 dla wynagrodzenia. Skorelowanie pomiędzy poszczególnymi zmiennymi objaśniającymi było jednak spore i np. dla współczynnika feminizacji i współczynnika urbanizacji wynosiło aż 0.82. Metodą Hellwiga wyznaczyliśmy najlepszy zestaw zmiennych objaśniających składający się z wynagrodzenia, współczynnika feminizacji, liczby oferty pracy i liczby małżeństw. Wykluczony został więc współczynnik urbanizacji. Można jednak przewidywać, że model ten nie będzie skutecznie dopasowywał się do zmiennych.

# DLA POWIATÓW GRODZKICH

## III – V. Wstępna analiza danych

Aby sprawdzić, jak dane prezentują się osobno dla powiatów grodzkich i powiatów ziemskich, podobną analizę przeprowadziłyśmy jeszcze dwa razy. Powiaty grodzkie (miasta na prawach powiatu) często różnią się od powiatów ziemskich pod względem struktury społeczno-gospodarczej, poziomu urbanizacji, dostępności do zasobów i usług oraz innych czynników, dlatego warto rozważyć stworzenie dla nich osobnych modeli. W tej części rozważymy dane tylko dla powiatów grodzkich i porównamy to z wynikami z poprzedniego przypadku.

Zdecydowaliśmy się jednak usunąć jedną z potencjalnych zmiennych objaśniających – współczynnik urbanizacji. Nie było sensu traktować go jako zmienną, gdyż w każdym przypadku dla powiatu grodzkiego (czyli miasta) wynosił on 100%, a odchylenie standardowe wynosiło 0.

## STATYSTYKI OPISOWE

BEZROBOCIE:

| Średnia | Odchylenie st. | Skośność | Kurtoza |
|---|---|---|---|
| 2.163816 | 0.9585595 | 0.9738661 | 0.2474233 |

Średnia bezrobocia wynosi około 2,16%. To oznacza, że na każde 100 osób około 2 osoby są bezrobotne. Jest to wartość o 0.6 p.p. mniejsza niż w przypadku uwzględnienia wszystkich powiatów. Oznacza to, że bezrobocie w miastach jest niższe niż bezrobocie ogółem.

Odchylenie standardowe wynosi około 0,96. Oznacza to, że od średniej wartości bezrobocia wartości w poszczególnych powiatach mogą różnić się o ok. 0.96 punktu procentowego. Odchylenie jest to niższe niż w poprzednim przypadku (1,33).

Skośność na poziomie około 0,97 wskazuje na prawostronną asymetrię rozkładu danych. Więcej powiatów ma niższe wskaźniki bezrobocia niż średnia. Poziom skośności jest nieznacznie niższy niż dla powiatów ogółem.

Kurtuza na poziomie 0,25 wskazuje na lekką leptokurtyczność. Jest to jednak wartość znacznie niższa niż uwzględniając wszystkie powiaty (0,95). Oznacza to, że rozkład bardziej przypomina rozkład normalny.

| Współczynnik zmienności | 44.29949 |
|---|---|

Wartość współczynnika zmienności na poziomie około 44,3% wskazuje na przeciętną, choć silną zmienność danych. Wartość ta jest nieznacznie niższa niż w poprzednich przypadku (48,3%).

WYNAGRODZENIA:

| Średnia | Odchylenie st. | Skośność | Kurtoza |
|---|---|---|---|
| 5780.596 | 775.8844 | 1.545652 | 3.105582 |

Średnie wynagrodzenie w miastach na prawach powiatu wynosi ok. 5780,6 złotych. Jest to wartość o ponad 550 złotych większa niż uwzględniając wszystkie powiaty.

Odchylenie standardowe wynosi ok. 775,9. Oznacza to, że od średniej wartości wynagrodzeń wartości wynagrodzeń w powiatach mogą różnić się o około 775,9 złotych. Wartość ta jest o ok. 135 złotych wyższa niż w poprzednim przypadku.

Skośność na poziomie 1,55 wskazuje na prawoskośność i oznacza, że jest więcej powiatów z zarobkami niższymi od średniej. Skośność jest jednak mniejsza niż w przypadku uwzględnienia wszystkich powiatów (2,66).

Kurtoza wynosi 3,1, co oznacza leptokurtyczność rozkładu. Wartość ta jest jednak zdecydowanie mniejsza niż dla wszystkich powiatów (12,4).

| Współczynnik zmienności | 13.42222 |
|---|---|

Współczynnik zmienności na poziomie 13,4% wskazuje na słabą zmienność danych, jednak wystarczającą w kontekście tworzenia modelu regresji. Wartość ta jest nieznacznie większa niż w poprzednim przypadku (12,3%).

WSPÓŁCZYNNIK FEMINIZACJI:

| Średnia | Odchylenie st. | Skośność | Kurtoza |
|---|---|---|---|
| 111.697 | 3.002408 | -0.0008348437 | -0.1444562 |

Średnia współczynnika feminizacji wynosi 111,7%. Oznacza to, że na każde 100 mężczyzn przypada średnio 111-112 kobiet. Jest to o ponad 6 kobiet więcej niż w powiatach ogółem, co oznacza, że proporcja kobiet do mężczyzn w miastach jest wyższa.

Odchylenie standardowe wynosi ok. 3. Oznacza to, że od średniej wartości współczynnika feminizacji wartości współczynnika feminizacji dla konkretnych powiatów mogą różnić się o około 3 punkty procentowe. Wartość ta jest nieznacznie mniejsza niż w poprzednim przypadku (3,99).

Skośność na poziomie -0.001 wskazuje na duże podobieństwo do rozkładu normalnego. W porównaniu do wartości dla wszystkich powiatów (1,05), która wskazywała na prawostronną asymetrię rozkładu, tutaj wartość statystyki jest nieznacznie ujemna,

Kurtoza również jest ujemna i względnie bliska zeru (-0.15). Wskazuje to na platykurtyczność, co oznacza, że wartości skupiają się bliżej średniej i prawdopodobnie nie występują wartości odstające. Dla powiatów ogółem ta statystyka miała wartość 0,66.

| | |
|---|---|
| Współczynnik zmienności | 2.687994 |

Wartość współczynnika zmienności na poziomie ok. 2,7% wskazuje na małą zmienność danych (nawet niższą niż w poprzednim przypadku – 3,8%) i jest zdecydowanie niezadawalająca w kontekście tworzenia modelu regresji: nie spełnia warunku V > 10%.

OFERTY PRACY:

| Średnia | Odchylenie st. | Skośność | Kurtoza |
|---|---|---|---|
| 424.0411 | 200.1777 | 0.8084019 | 0.07115545 |

Średnia liczba ofert pracy w ciągu roku wynosi około 424,04. Oznacza to, że przeciętnie na 10 tysięcy mieszkańców miasta na prawach powiatu przypadają około 424 oferty pracy. Jest to wartość zdecydowanie wyższa niż w przypadku uwzględnienia wszystkich powiatów – wtedy na 10 tys. mieszkańców przypadały średnio 343 oferty.

Odchylenie standardowe wynosi ok. 200,18. Oznacza to, że od średniej wartości liczby ofert pracy (na 10 tys. mieszkańców) wartości liczby ofert pracy w powiatach mogą różnić się o około 200 ofert.

Skośność wynosi 0,81, co oznacza prawostronną asymetrię rozkładu. Wartość ta jest jednak niższa niż w poprzednim przypadku (1,16).

Kurtoza wynosi ok. 0,07. Jest to wartość bliska zeru i sugeruje względne podobieństwo rozkładu do rozkładu normalnego. W przypadku uwzględnienia wszystkich powiatów kurtoza wynosiła 3,06, co oznaczało większe zagęszczenie wartości wokół średniej.

| | |
|---|---|
| Współczynnik zmienności | 47.20716 |

Wartość współczynnika zmienności wynosi ok. 42%. Wskazuje to na przeciętną/silną zmienność danych, choć wartość ta jest nieco niższa niż przy uwzględnieniu wszystkich powiatów (56%). Wciąż spełnia jednak warunek V > 10%.

LICZBA MAŁŻEŃSTW:

| Średnia | Odchylenie st. | Skośność | Kurtoza |
|---|---|---|---|
| 43.83299 | 4.850288 | 0.6348509 | 1.117276 |

Średnia liczba zawartych małżeństw wynosi ok. 43,83. Oznacza to, że w miastach na prawach powiatu średnio na każde 10 tysięcy mieszkańców przypadają 43-44 zawarte małżeństwa. Nie widać tu zbytniej różnicy między powiatami grodzkimi a ziemskimi.

Odchylenie standardowe wynosi około 4,85. Oznacza to, że od średniej wartości liczby zawartych małżeństw (43,83) wartości liczby małżeństw w powiatach grodzkich mogą różnić się o około 4,85. Wartość ta nie różni się znacznie od wyznaczonej dla powiatów ogółem (4,56).

Skośność wynosi 0,63, co wskazuje na prawostronną asymetrię rozkładu. Sugeruje to, że istnieje tendencja do częstszych wystąpień niższych wartości, ale wartości te nie odbiegają znacząco od średniej. Statystyka ta wynosi jednak więcej niż w przypadku uwzględnienia wszystkich powiatów (0,13).

Kurtuza wynosi ok. 1,12 i mówi o leptokurtyczności rozkładu. Wartość ta jest wyższa niż w poprzednim przypadku (0,24).

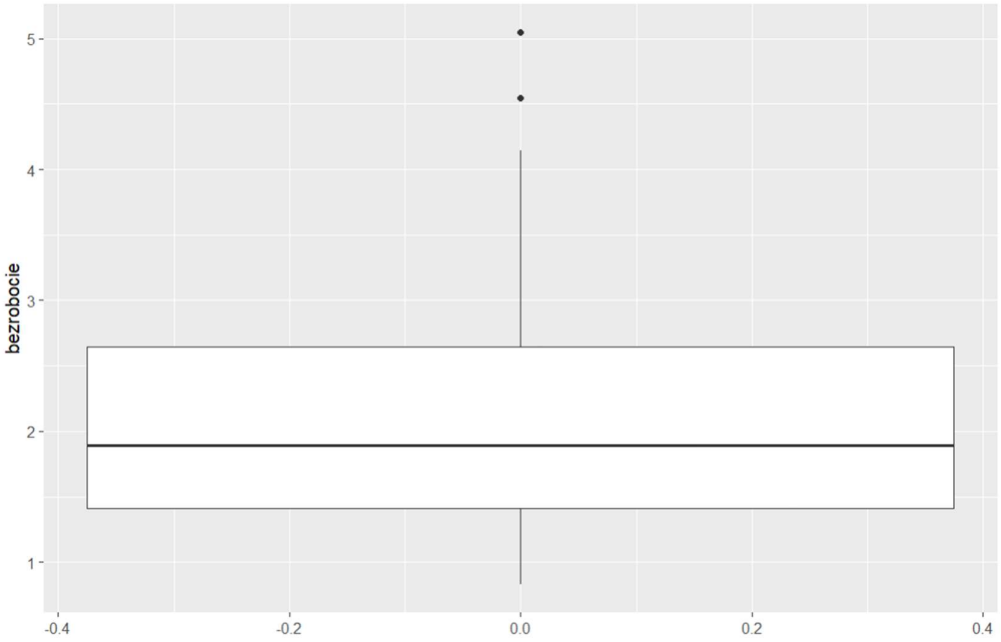
| Współczynnik zmienności | 11.06538 |
|---|---|

Współczynnik zmienności wynosi ok. 11,07%. Jest to wartość nieznacznie wyższa niż dla powiatów ogółem (10,55%) i spełnia warunek V > 10%.

## WYKRESY PUDEŁKOWE

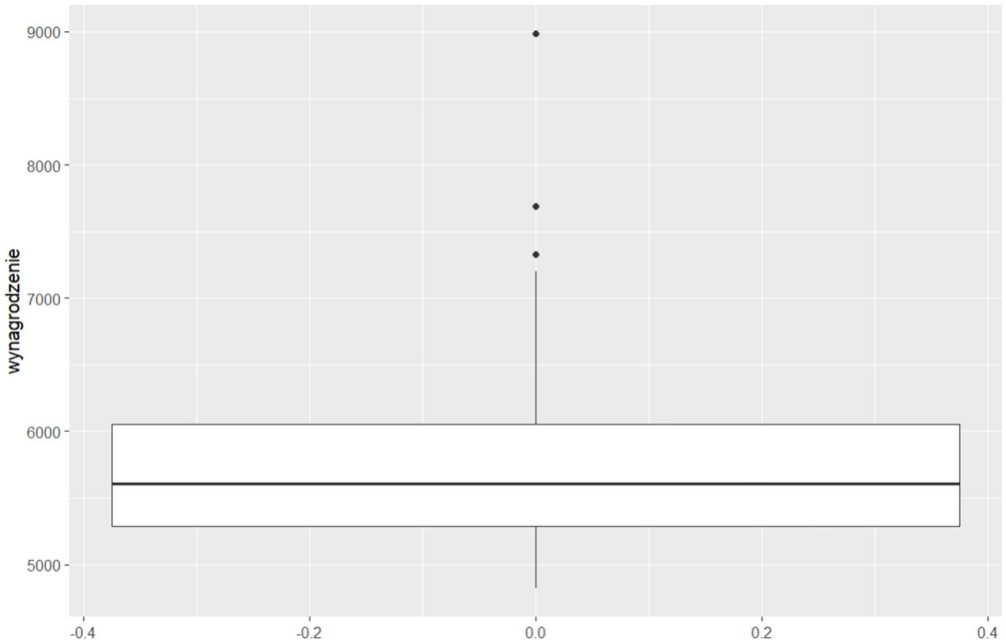
Aby lepiej zobrazować rozkłady danych, a przede wszystkich ich wartości odstające, ponownie narysowałyśmy wykresy pudełkowe.

BEZROBOCIE:



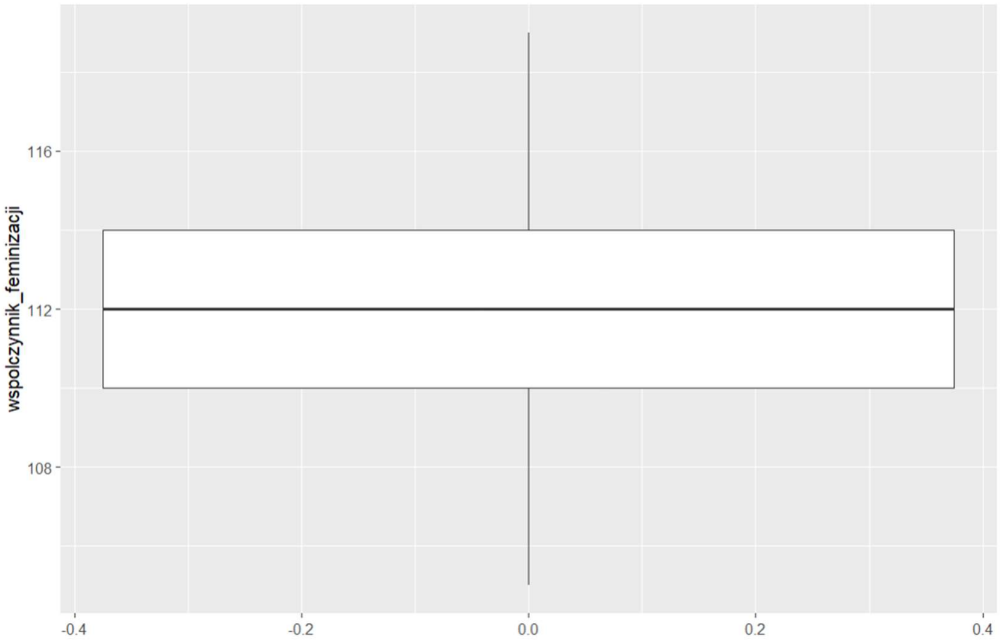

Rozkład bezrobocia jest prawostronnie skośny i posiada parę wysokich wartości odstających. Górny wąs wykresu jest dłuższy niż dolny.

WYNAGRODZENIE:



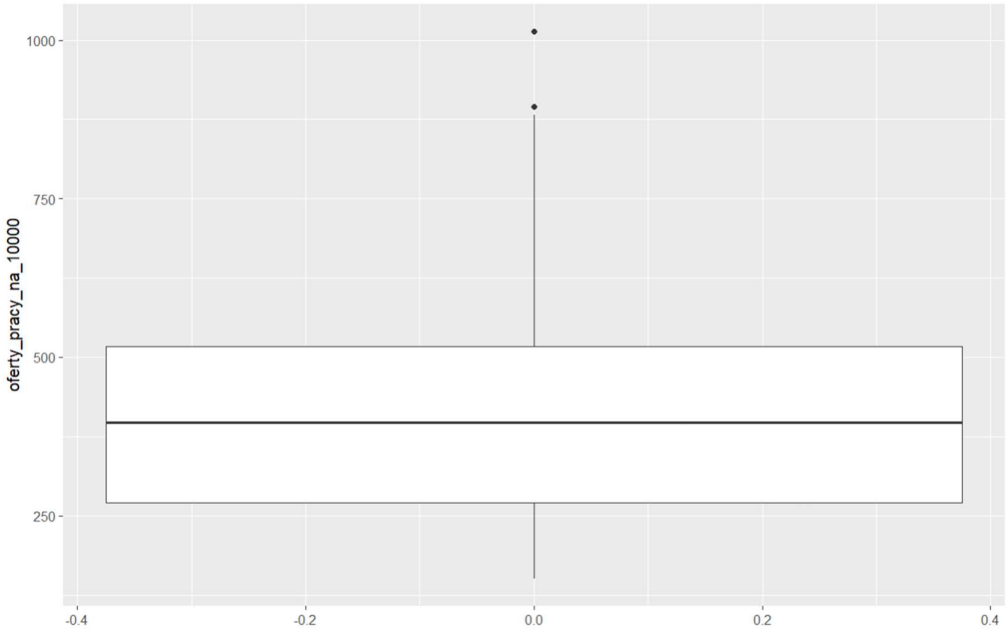
Rozkład wynagrodzenia również jest prawoskośny i posiada pojedyncze, wysokie outliery.

WSPÓŁCZYNNIK FEMINIZACJI:



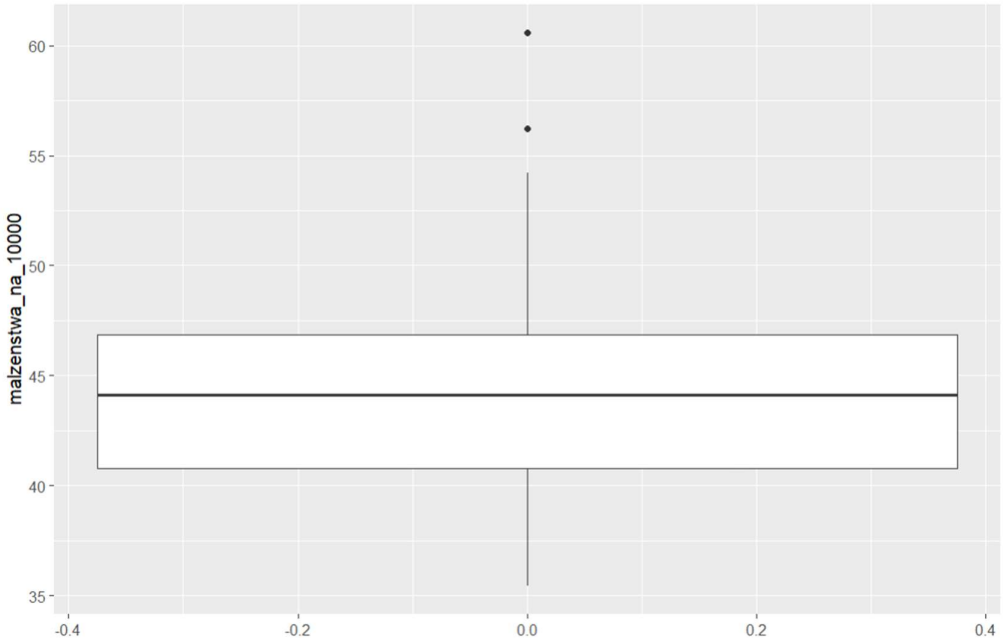

Wykres współczynnika feminizacji jest symetryczny i może sugerować podobieństwo do rozkładu normalnego. Nie widać żadnych wartości odstających.

OFERTY PRACY:




Rozkład ofert pracy w ciągu roku jest prawostronnie skośny i posiada pojedyncze outliery.

LICZBA MAŁŻEŃSTW:




Wykres liczby małżeństw również jest prawostronnie asymetryczny i posiada niewiele wartości odstających.

W tym przypadku również zdecydowaliśmy się na nieusuwanie wartości odstających, aby model brał pod uwagę wszystkie rzeczywiste wartości zebrane w powiatach.

# VI. KORELACJE

Następnie policzyłyśmy korelacje między wszystkimi zmiennymi.

| | Bezrobocie | Wynagrodzenie | Ws. feminizacji | Oferty pracy | Małżeństwa |
|---|---|---|---|---|---|
| Bezrobocie | 1.0 | -0.2480 | 0.1662 | -0.1846 | -0.3322 |
| Wynagrodzenie | -0.2480 | 1.0 | 0.1317 | -0.0756 | 0.4037 |
| Ws. feminizacji | 0.1662 | 0.1317 | 1.0 | 0.0323 | -0.1421 |
| Oferty pracy | -0.1846 | -0.0756 | 0.0323 | 1.0 | 0.0274 |
| Małżeństwa | -0.3322 | 0.4037 | -0.1421 | 0.0274 | 1.0 |

Jak widać, bezrobocie (zmienna objaśniana) nie posiada wysokiej korelacji z żadną ze zmiennych objaśniających. Najwyższą wartość, -0.3322, osiąga dla liczby zawieranych małżeństw. Pozostałe wartości (-0.2480 dla przeciętnego wynagrodzenia, -0.1846 dla liczby ofert pracy i 0.1662 dla współczynnika feminizacji) są tym bardziej niezadawalające w kontekście tworzenia modelu regresji. Z powodu usunięcia współczynnika urbanizacji ze zmiennych objaśniających nie występuje pomiędzy nimi aż tak silna korelacja, jednak wartość 0.4037 dla pary wynagrodzenie – liczba małżeństw wciąż jest niepokojąca.

# VII. METODA HELLWIGA

Ponownie, aby wyznaczyć najlepszy zestaw zmiennych objaśniających użyłyśmy metody Hellwiga.

```
Największa wartość integralnej pojemności informacyjnej: 0.1595319
Najlepszy podzbiór zmiennych: X1  X2  X3  X4
```

Jak widać, program nie wyeliminował żadnej zmiennej i uznał, że poziom bezrobocia w powiatach grodzkich najlepiej tłumaczy poziom przeciętnego wynagrodzenia, współczynnik feminizacji, liczba ofert pracy w ciągu roku i liczba zawieranych małżeństw.

## WNIOSKI

Rozkłady części zmiennych objaśniających różnią się od tych zawierających dane ze wszystkich powiatów. Średnie wynagrodzenie, współczynnik feminizacji i liczba ofert pracy są wyższe w powiatach grodzkich niż w tych ziemskich. Rozkład współczynnika feminizacji jako jedyny charakteryzował się (nieznacznie) ujemną skośnością i kurtozą. Poziom jego współczynnika zmienności wynosił 2,7%, co jest niezadawalającą wartością w kontekście tworzenia modelu regresji. Za pomocą metody Hellwiga wyznaczyliśmy najlepszy zestaw zmiennych objaśniających, czyli poziom przeciętnego wynagrodzenia, współczynnik feminizacji, liczbę ofert pracy w ciągu roku i liczbę zawieranych małżeństw. Jest to sam zbiór co ten wyznaczony dla danych ze wszystkich powiatów.

# DLA POWIATÓW ZIEMSKICH

## III – V. Wstępna analiza danych

Następnie wybrałyśmy jedynie powiaty ziemskie, aby zobaczyć, jak dane różnią się od tych wyznaczonych dla powiatów grodzkich i wszystkich powiatów ogółem. W tym przypadku ponownie uwzględniłyśmy współczynnik urbanizacji.

### STATYSTYKI OPISOWE

BEZROBOCIE:

| Średnia | Odchylenie st. | Skośność | Kurtoza |
|---|---|---|---|
| 2.873815 | 1.363064 | 0.9584006 | 0.7543345 |

Średnie bezrobocie w powiatach ziemskich wynosi około 2.9%. Jest to nieznacznie więcej niż we wszystkich powiatach ogółem.

Odchylenie standardowe bezrobocia to około 1.36. Mówi to o tym, że bezrobocie może różnić się od średniej o około 1.33 punkty procentowe.

Skośność wynosi około 0.95. Oznacza to, że rozkład jest prawostronnie skośny, a więc większość powiatów ma bezrobocie poniżej średniej, czyli poniżej 2.9%.

Kurtoza bezrobocia jest dodatnia i wynosi około 0.75. Sugeruje to, że mogą pojawić się wartości odstające.

| Współczynnik zmienności | 47.43046 |
|---|---|

Współczynnik zmienności bezrobocia wynosi 47%. Jest to przeciętna zmienność, jednak zbliżająca się już do silnej. Wartość ta jest bardzo podobna do wartości współczynnika zmienności dla wszystkich powiatów ogółem.

WYNAGRODZENIA:

| Średnia | Odchylenie st. | Skośność | Kurtoza |
|---|---|---|---|

| | | | |
|---|---|---|---|
| 5093.955 | 537.5683 | 3.534193 | 24.50747 |

Średnie wynagrodzenie wynosi około 5093.96. Oznacza to, że przeciętne wynagrodzenie mieszkańca powiatu ziemskiego to właśnie około 5093 zł.

Odchylenie standardowe wynagrodzenia dla powiatów ziemskich to około 537 zł. Jest to prawie 100 zł mniej niż w przypadku danych uwzględniających wszystkie powiaty.

Skośność na poziomie 3.53 wskazuje na fakt, że większość powiatów znajduje się jednak poniżej średniej, czyli w większości powiatów ludzie średnio zarabiają mniej niż 5093 zł.

Kurtoza dla powiatów ziemskich jest ponad 10 większa niż dla wszystkich powiatów i wynosi aż 24.5. Jest to rozkład leptokurtyczny. Wartość kurtozy sugeruje duże skoncentrowanie danych wokół średniej i duże prawdopodobieństwo wystąpienia wartości odstających.

| | |
|---|---|
| Współczynnik zmienności | 10.55306 |

Współczynnik zmienności nie różni się znacząco od zmienności dla wszystkich powiatów. Jest on jeszcze niższy niż tamten, jednak w dalszym ciągu przekracza poziom 10 procent.

WSPÓŁCZYNNIK FEMINIZACJI:

| Średnia | Odchylenie st. | Skośność | Kurtoza |
|---|---|---|---|
| 103.5897 | 2.428056 | 0.5109896 | 0.2713614 |

Średnia dla współczynnika feminizacji wynosi ok. 103,6. Oznacza to, że w powiatach ziemskich na 100 mężczyzn przypadają około 103 kobiety.

Odchylenie standardowe to około 2.42. Jest to mniej niż w przypadku danych dla wszystkich powiatów. Oznacza to, że wartości współczynnika feminizacji dla konkretnych powiatów ziemskich mogą różnić się o około 2.42 punkty procentowe

Skośność wynosi ok. 0.5, więc w dalszym ciągu sugeruje prawostronność rozkładu. Jest ona jednak mniejsza niż w przypadku danych dla wszystkich powiatów.

Kurtoza współczynnika feminizacji jest mniejsza niż w przypadku pierwszych danych i wynosi około 0.27.

| | |
|---|---|
| Współczynnik zmienności | 2.343916 |

Współczynnik zmienności również jest mniejszy i wynosi 2.34%, a więc ponownie nie spełnia on warunku mówiącego o tym, że zmienność danych objaśniających w modelu powinna być na poziomie minimum 10%.

WSPÓŁCZYNNIK URBANIZACJI:

| Średnia | Odchylenie st. | Skośność | Kurtoza |
|---|---|---|---|
| 40.25385 | 16.60179 | -0.02527276 | 0.7543345 |

Współczynnik urbanizacji dla powiatów ziemskich wynosi średnio 40.25%. Oznacza to, że ponad połowa ludzi zamieszkuje tereny wiejskie.

Odchylenie standardowe to natomiast około 16 punktów procentowych. Jest to o prawie 10 punktów procentowych mniej niż w przypadku danych dla wszystkich powiatów.

Ujemna skośność powinna wskazywać na fakt, że współczynnik ten dla większości powiatów znajduje się powyżej średniej. Jednak w tym przypadku wynosi on jedynie 0.0252, a więc można stwierdzić, że rozkład ten jest stosunkowo symetryczny.

Kurtoza również nie różni się wiele od zera (0.75), jednak jest ona dodatnia i sugeruje leptokurtyczność rozkładu.

| | |
|---|---|
| Współczynnik zmienności | 41.24273 |

Współczynnik zmienności wynosi 41,24%. Wskazuje to na całkiem silną zmienność i zróżnicowanie danych, jednak jest to mniej niż w przypadku danych dla wszystkich powiatów.

OFERTY PRACY:

| Średnia | Odchylenie st. | Skośność | Kurtoza |
|---|---|---|---|
| 326.4343 | 186.6711 | 1.878718 | 4.432094 |

Średnia liczba ofert pracy w powiatach ziemskich to około 326. Wartość ta nie różni się znacznie od danych dla wszystkich powiatów (343 oferty), jednak jest odrobinę niższa.

W odchyleniu standardowym również nie ma znaczącej różnicy. Wynosi ono około 186.67 i jest mniejsze o około 6 ofert.

Zarówno skośność, jak i kurtoza mają wartości podobne do danych dla wszystkich powiatów. Skośność wynosi bowiem 1.88, natomiast kurtoza 4.43.

| Współczynnik zmienności | 57.18488 |
|---|---|

Współczynnik zmienności na poziomie 57%, podobnie jak w danych dla wszystkich powiatów, wskazuje na silną zmienność danych i spełnia warunek V > 10%.

LICZBA MAŁŻEŃSTW:

| Średnia | Odchylenie st. | Skośność | Kurtoza |
|---|---|---|---|
| 43.13842 | 4.500821 | -0.01798265 | -0.1411562 |

Zarówno średnia liczba małżeństw, jak i odchylenie standardowe są na tych samym poziomie co w przypadku wszystkich powiatów, a więc wynoszą kolejno 43 i 4.5.

Skośność natomiast różni się nieznacznie, gdyż jest ona ujemna. Wartość ta jest natomiast na tyle mała, iż można stwierdzić, że rozkład ten jest symetryczny i pod tym aspektem podobny do normalnego.

Wartość kurtozy również jest niewielka (-0.14). Dane więc przypominają rozkład normalny.

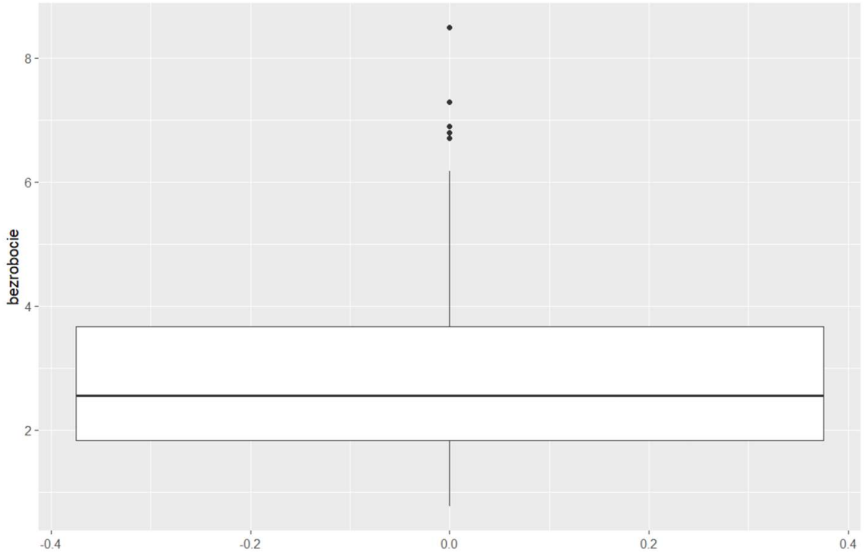
| Współczynnik zmienności | 10.43344 |
|---|---|

Współczynnik zmienności na poziomie 10% to podobnie jak w przypadku danych dla wszystkich powiatów niewiele więcej niż minimalna granica, jednak jest on wciąż akceptowalny.

## WYKESY PUDEŁKOWE

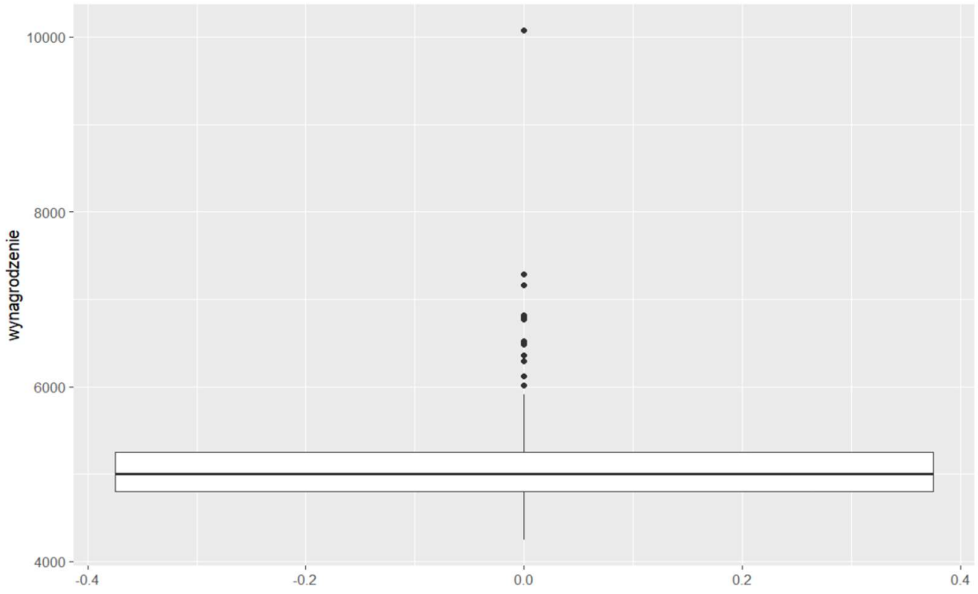
Ponownie dla zoobrazowania danych przedstawiłyśmy wykresy pudełkowe dla rozkładów poszczególnych zmiennych.

### BEZROBOCIE



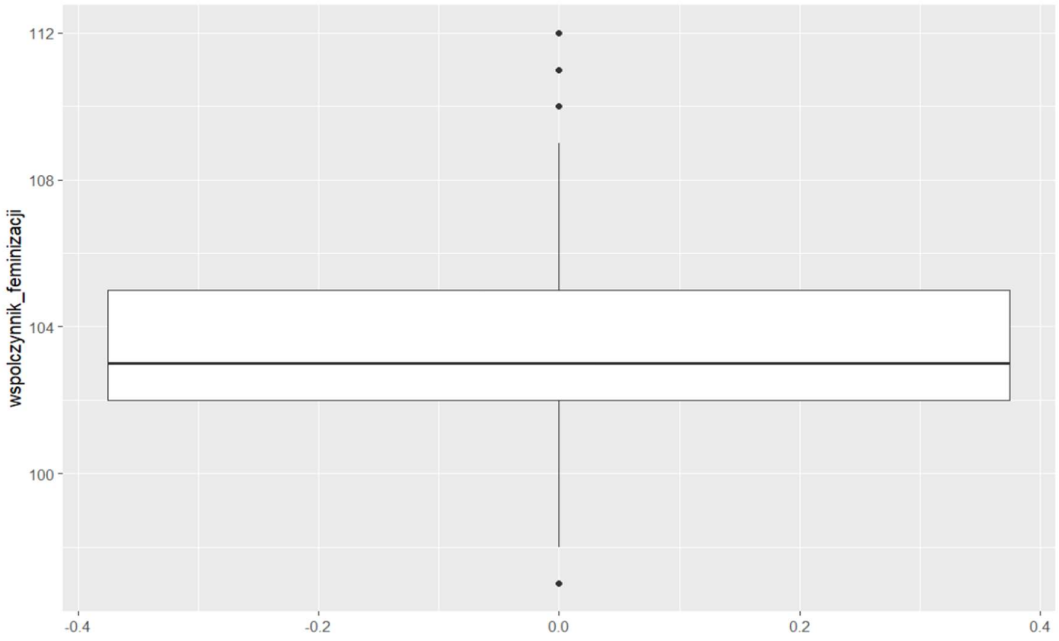

Jak można zauważyć, rozkład jest zdecydowanie prawostronnie skośny i posiada kilka wartości odstających.

WYNAGRODZENIE



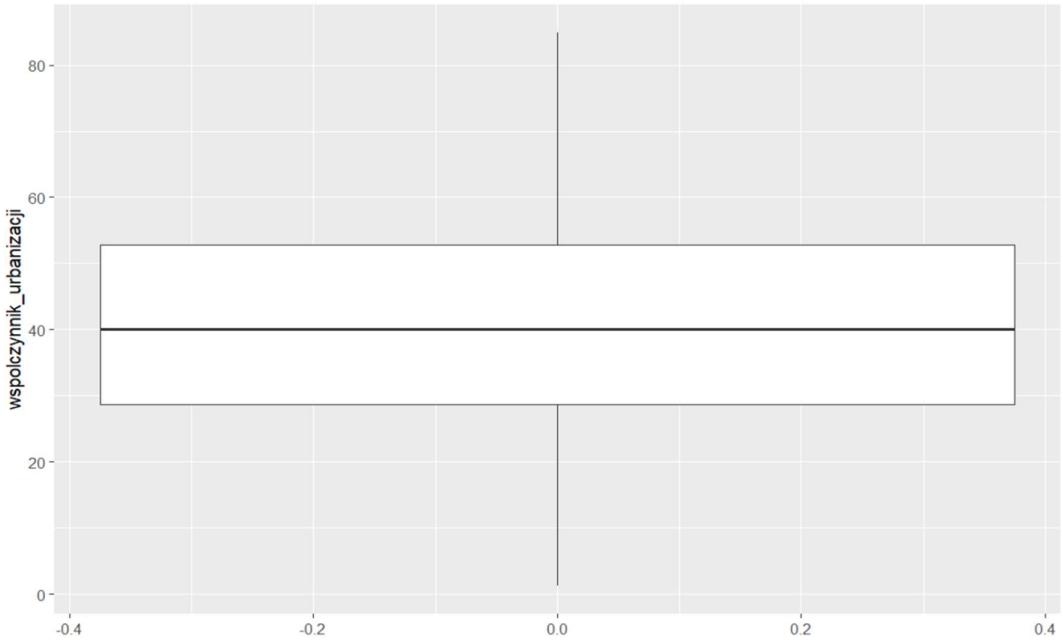

Można zauważyć, że wykres pokrywa się z obliczonymi statystykami i występuje spora ilość wartości odstających.

WSPÓŁCZYNNIK FEMINIZACJI



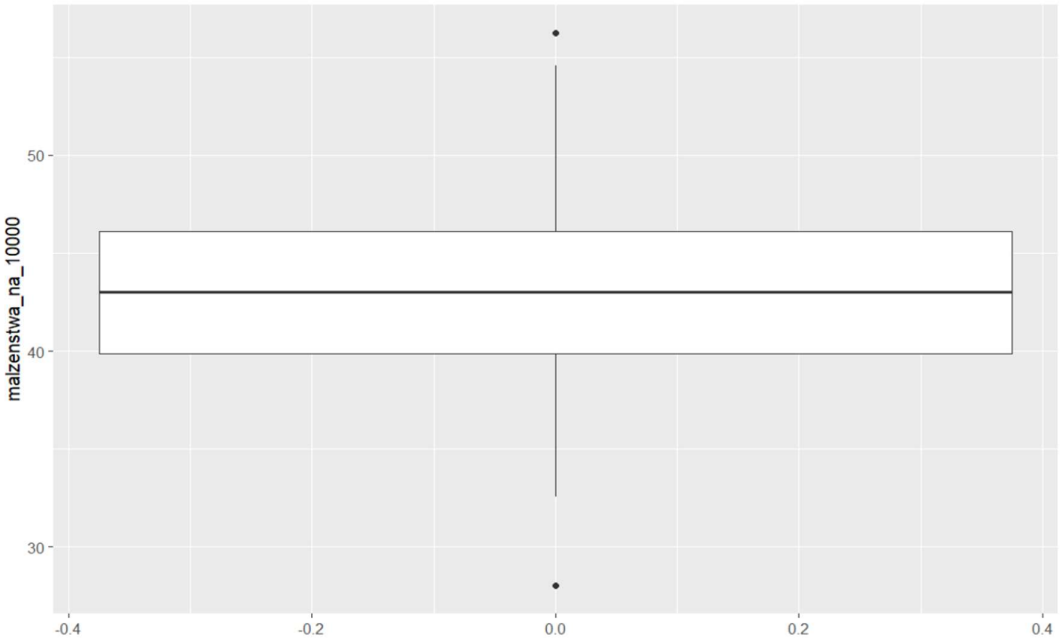
Wykres jest widocznie prawoskośny i posiada parę wartości odstających, czego nie można było zauważyć podczas analizowania obliczonych statystyk.

WSPÓŁCZYNNIK URBANIZACJI




Wykres jest symetryczny i przypomina rozkład normalny. Nie można zaobserwować żadnych wartości odstających.

LICZBA MAŁŻEŃSTW



Rozkład posiada niewiele wartości odstających. Wykres przypomina rozkład normalny, co pokrywa się z obliczonymi statystykami.

# VI. KORELACJE

| | Bezrobocie | Wynagrodzenie | Ws. feminizacji | Ws. urbanizacji | Oferty pracy | Małżeństwa |
|---|---|---|---|---|---|---|
| Bezrobocie | 1.0 | -0.2234 | -0.2331 | -0.1531 | -0.1573 | -0.0525 |
| Wynagrodzenie | -0.2234 | 1.0 | 0.3182 | 0.2592 | 0.2082 | -0.0440 |
| Ws. feminizacji | -0.2331 | 0.3182 | 1.0 | 0.5852 | 0.1869 | -0.3160 |
| Ws. urbanizacji | -0.1531 | 0.2592 | 0.5852 | 1.0 | 0.1921 | -0.3283 |
| Oferty pracy | -0.1573 | 0.2082 | 0.1869 | 0.1921 | 1.0 | -0.0948 |
| Małżeństwa | -0.0525 | -0.0440 | -0.3160 | -0.3283 | -0.0948 | 1.0 |

Obliczone korelacje dla powiatów ziemskich różnią się od korelacji tych zmiennych dla wszystkich powiatów, jednak różnice te nie są aż tak znaczące. W tym przypadku zmienna objaśniająca najsilniej skorelowana ze zmienną objaśnianą to współczynnik feminizacji, podczas gdy wcześniej było to wynagrodzenie. Różnice w wartościach są jednak bardzo niewielkie. Korelacja współczynnika feminizacji z bezrobociem to -0.23, natomiast wynagrodzenia z bezrobociem to –0.22. Korelacja pozostałych zmiennych objaśnianych z bezrobociem przedstawia się następująco:

- Współczynnik urbanizacji: -0.15,
- Oferty pracy: -0.15,
- Małżeństwa: -0.05.

Podobnie jak w przypadku danych dla wszystkich powiatów korelacje te nie są korzystne dla modelu. W przypadku korelacji między zmiennymi objaśniającymi wartości wahają się przeważnie między 0.1 a 0.3, czyli są względnie niskie. Największą korelację zaobserwować można pomiędzy współczynnikiem feminizacji i urbanizacji. Wynosi ona 0.58 i choć jest

niższa niż w przypadku korelacji tych samych zmiennych dla wszystkich powiatów, wciąż jest niekorzystna dla modelu.

## VII. METODA HELLWIGA

Za pomocą metody Hellwiga wyznaczyłyśmy najlepszy zestaw zmiennych objaśniających.

> Największa wartość integralnej pojemności informacyjnej: 0.08656858
> Najlepszy podzbiór zmiennych: X1  X4  X5

Jako najlepszy podzbiór zmiennych objaśniających poziom bezrobocia wyznaczone zostało przeciętne wynagrodzenie, liczba ofert pracy w ciągu roku i liczba zawieranych małżeństw. Program odrzucił współczynniki feminizacji oraz urbanizacji.

### WNIOSKI

Analiza statystyk opisowych dla powiatów ziemskich pokazała, że powiaty te charakteryzują się wyższym bezrobociem niż powiaty grodzkie, a także niższym przeciętnym wygrodzeniem, współczynnikiem feminizacji oraz liczbą ofert pracy w ciągu roku. Rozkłady bezrobocia i przeciętnego wynagrodzenia posiadają sporą ilość wartości odstających, jednak ponownie zdecydowałyśmy się je zachować dla przedstawienia rzeczywistego obrazu zjawiska. Zmienne objaśniające nie posiadają silnej korelacji ze zmienną objaśnianą, co wpłynie negatywnie na model regresji. Problemem był również wysoki poziom korelacji między wskaźnikiem urbanizacji a wskaźnikiem feminizacji, jednak za pomocą metody Hellwiga odrzuciliśmy te zmienne. Najlepszym zestawem zmiennych objaśniających poziom bezrobocia w powiatach ziemskich jest przeciętne wynagrodzenie, liczba ofert pracy w ciągu roku i liczba zawieranych małżeństw.

# PODSUMOWANIE

Wybrane przez nas dane, czyli średnie wynagrodzenie, współczynnik feminizacji, współczynnik urbanizacji, liczba ofert pracy w ciągu roku i liczba małżeństw zawieranych w ciągu roku różnią się między powiatami grodzkimi i ziemskimi. W powiatach grodzkich można zaobserwować nieznacznie niższy poziom bezrobocia, a także wyższy poziom

średniego wynagrodzenia, wskaźnika feminizacji i ofert pracy. Liczba zawieranych małżeństw w ciągu roku zdaje się nie zależeć od rodzaju powiatu. We wszystkich wariantach (dla wszystkich powiatów, tylko grodzkich i tylko ziemskich) korelacje zmiennej objaśnianej ze zmiennymi objaśniającymi są niezbyt wysokie; występują natomiast wysokie i umiarkowane korelacje między zmiennymi objaśniającymi. W każdym przypadku wskaźnik feminizacji ma zbyt niski współczynnik zmienności, by uwzględnić go w modelu regresji. Wykluczony został również wskaźnik urbanizacji. Za pomocą metody Hellwiga udało nam się wyznaczyć najlepsze zestawy zmiennych objaśniających: dla wszystkich powiatów ogółem jest to przeciętne wynagrodzenie, współczynnik feminizacji (charakteryzujący się zbyt niskim poziomem zmienności), liczba ofert pracy i liczba zawieranych małżeństw. Dla powiatów grodzkich najlepszy podzbiór zmiennych się nie zmienił, a dla powiatów ziemskich odrzucony został dodatkowo współczynnik feminizacji.