Mateusz Stasiak

Analiza danych rzeczywistych

1. Wstęp

Celem ćwiczenia jest analiza, porównanie i wizualizacja danych rzeczywistych przy pomocy metod statystyki opisowej.

2. Pierwszy zestaw danych

2.1. Wprowadzenie

Dane zostały pozyskane ze strony internetowej `https://vincentarelbundock.github.io/Rdatasets/datasets.html`. Dotyczą zarobków w Stanach Zjednoczonych w latach 1988 − 1989. Składają się z 4266 obserwacji. Każdej z nich przypisano roczną pensję w dolarach oraz odpowiednią grupę wiekową:
— g1 - od 23 do 26 lat,
— g2 - od 27 do 29 lat,
— g3 - od 30 do 32 lat.

	age	y
1	g3	569.5
2	g3	895.5
3	g3	1111.0
4	g3	1182.0
5	g3	1277.5
6	g3	1384.0

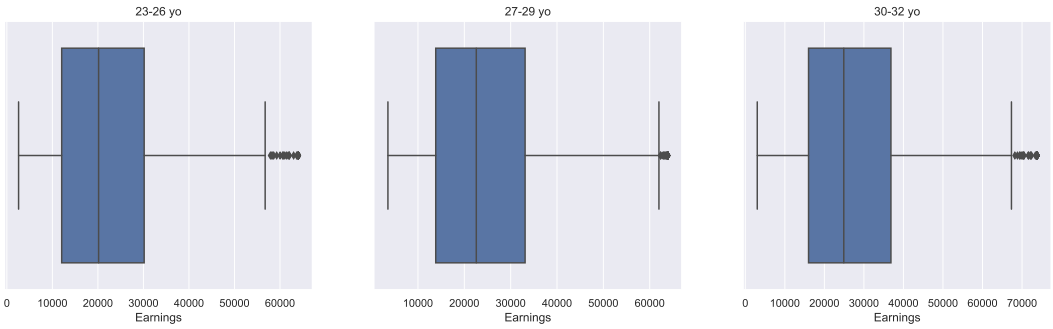
Rysunek 1. Przykładowe dane

2.2. Analiza danych

Efektywne opracowanie powyższych danych wymaga rozpatrywania każdej grupy wiekowej osobno.

	Earnings in group 1	Earnings in group 2	Earnings in group 3
count	1109.000000	1678.000000	1479.000000
mean	22880.475654	25080.174017	27973.634550
std	14667.586946	15291.887035	16502.686426
min	332.500000	886.500000	569.500000
25%	12040.500000	13831.875000	15995.500000
50%	20151.000000	22584.000000	24909.500000
75%	30161.500000	33115.375000	36838.750000
max	83810.500000	83810.500000	83810.500000
harmonic mean	13344.302964	15522.724505	16880.419764
geometric mean	18364.198964	20461.678884	22960.326130

Rysunek 2. Statystyki opisowe dla trzech grup wiekowych



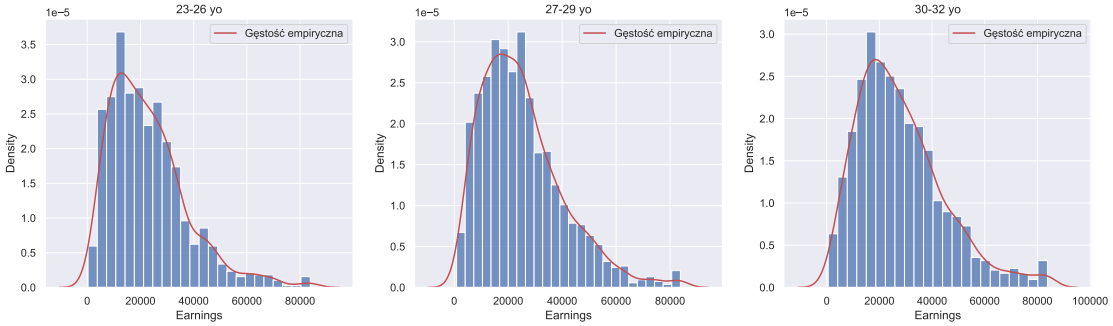
Rysunek 3. Wykresy pudełkowe dla trzech grup wiekowych

Wartości średnie wynagrodzeń rosną wraz z numerem grupy, który jest warunkowany wiekiem. Może to mieć związek z większym doświadczeniem pracownika. Bardzo wysokie odchylenie standardowe w stosunku do średniej (około 60%) oznacza, że w każdej grupie wiekowej są ogromne dysproporcje zarobków. Porównując wartości minimalne i maksymalne otrzymujemy różnicę rzędu 80 tysięcy. Uwzględniając fakt, że średnie są zawsze większe od percentyla 50% o około 2,5 tysiąca (12%), można wywnioskować, że rozbieżność płac jest spowodowana jednostkami, które zarabiają znacznie więcej od pozostałych. Zgadza się to z wykresami pudełkowymi. Na każdym z nich można zaobserwować wartości odstające z prawej strony. Co więcej, owe wykresy zostały sporządzone z danych po winsoryzacji. Zastąpienie skrajnych danych najbliższymi wartościami spowodowało wyraźne spadki wartości średnich zarobków, zwłaszcza w pierwszej grupie. Świadczy to o nielicznych relatywnie wysokich obserwacjach, które zawyżają statystyki opisowe.

Dane są zgodne z zasadą Pareto. Sumując ze sobą 20% największych wynagrodzeń, a następnie dzieląc je przez zarobki wszystkich pracowników otrzymamy 77%, czyli wartość zbliżoną do progu 80%.

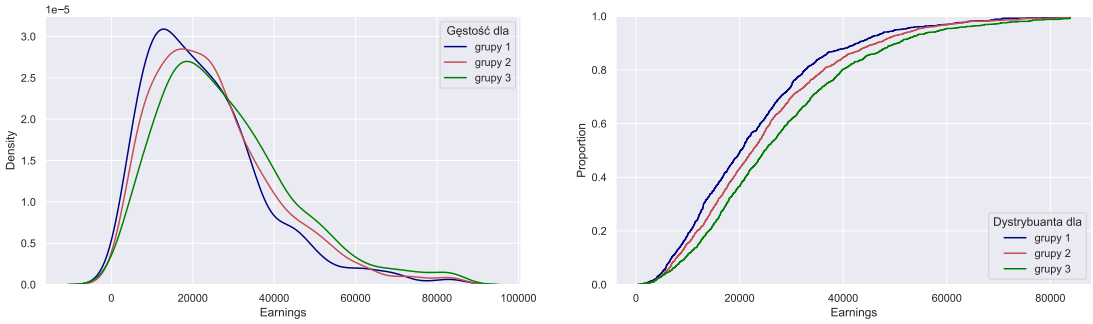
Dla porównania, średnie zarobki w Polsce na przełomie lat 1988/1989 wynosiły około 480 dolarów rocznie, czyli ponad 50 razy mniej.

2.3. Analiza rozkładu danych



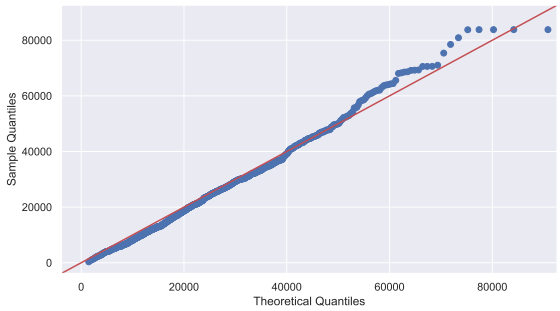
Rysunek 4. Histogramy dla trzech grup wiekowych

Histogramy nie oddają w pełni różnic otrzymanych w tabeli. Mają zbliżone kształty, a wcześniej wykazane dysproporcje obrazują ostatnie słupki. Najwięcej pracowników zarabiających w okolicy 80 tysięcy należy do najstarszej grupy wiekowej, co powoduje zawyżenie statystyk opisowych. Istotną różnicą są wyraźne dominanty w pierwszej i trzeciej grupie, podczas gdy środkowy wykres jest bardziej wypłaszczony. Może to świadczyć o istnieniu standardowych pensji dla pracowników w przedziałach wiekowych 23-26 i 30-32 lat. Innym wyjaśnieniem tego zjawiska może być proces nabywania doświadczenia. W drugiej grupie wiekowej znajdują się osoby, które otrzymały już podwyżkę oraz osoby, które nadal są w trakcie szkolenia i mają awans przed sobą.



Rysunek 5. Porównanie gęstości i dystrybuant dla trzech grup wiekowych

Zbliżone kształty wykresów sugerują, że dane z trzech grup wiekowych mają ten sam rozkład, ale z różnymi parametrami. Wygląd krzywych wskazuje na rozkład gamma z dwoma parametrami.



Rysunek 6. Wykres kwantylowy dla pierwszej grupy

Wykres kwantylowy został wykonany z argumentem Gamma(3,0,8000). Dane układają się wzdłuż czerwonej linii nachylonej do osi OX pod kątem 45°, zatem rozkład zarobków pierwszej grupy wiekowej jest zbliżony do rozkładu Gamma(3,8000).

3. Drugi zestaw danych

3.1. Wprowadzenie

Dane zostały pozyskane ze strony internetowej `https://vincentarelbundock.github.io/Rdatasets/datasets.html`. Dotyczą zarobków w Stanach Zjednoczonych w 1982 roku. Składają się z 595 obserwacji. Każdej z nich przypisano 12 zmiennych, w tym miesięczną pensję w dolarach oraz informację o stanie małżeńskim pracownika.

	experience	weeks	occupation	industry	south	smsa	married	gender	union	education	ethnicity	wage
1	9	32	white	yes	yes	no	yes	male	no	9	other	515
2	36	30	blue	yes	no	no	yes	male	no	11	other	912
3	12	46	blue	yes	no	no	no	male	yes	12	other	954
4	37	46	blue	no	no	yes	no	female	no	10	afam	751
5	16	49	white	no	no	no	yes	male	no	16	other	1474
6	32	47	blue	yes	no	yes	yes	male	no	12	other	1539

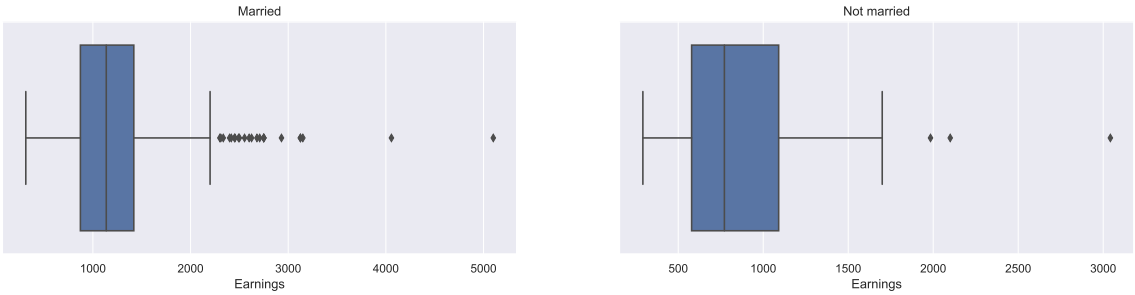
Rysunek 7. Przykładowe dane

3.2. Analiza danych

Efektywne badanie korelacji między pensją a stanem małżeńskim wymaga rozpatrywania obu tych grup osobno.

	Married	Not married
count	479.000000	116.000000
mean	1214.745303	872.586207
std	532.054738	430.268027
min	313.000000	292.000000
25%	872.000000	578.750000
50%	1137.000000	771.500000
75%	1419.000000	1090.750000
max	5100.000000	3042.000000
harmonic mean	1031.586166	706.029893
geometric mean	1118.915372	783.907880

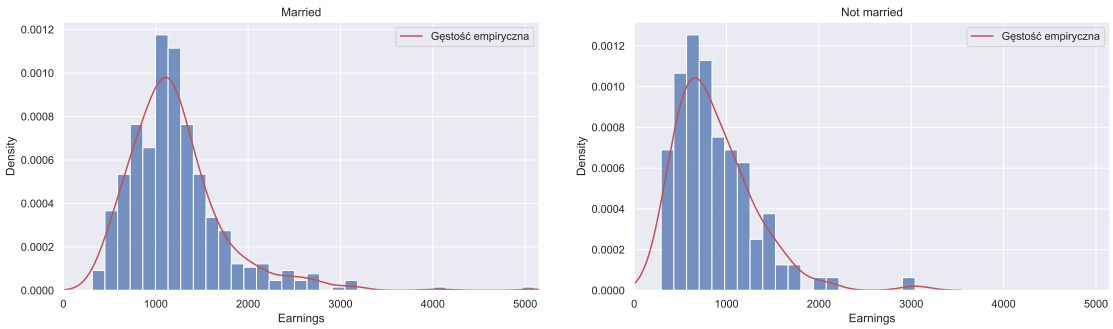
Rysunek 8. Statystyki opisowe dla obu stanów cywilnych



Rysunek 9. Wykres pudełkowy dla obu stanów cywilnych

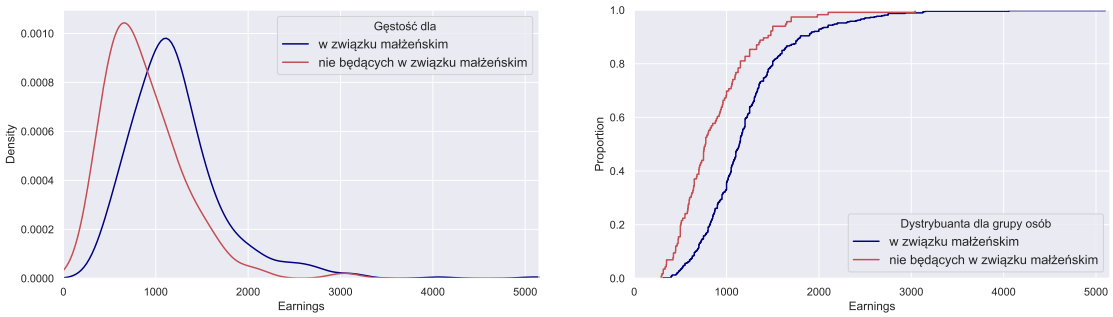
Zdecydowaną większość obserwacji stanowią osoby w związku małżeńskim. Średnie wynagrodzenie dla tej grupy osób jest zdecydowanie wyższe. Może to mieć związek z wiekiem, a co za tym idzie, z większym doświadczeniem zawodowym. Innym powodem może być po prostu chęć zabezpieczenia rodziny finansowo. Ponownie możemy zaobserwować bardzo wysokie odchylenie standardowe w stosunku do średniej (43% i 49%). To oznacza, że w obu grupach występują dysproporcje zarobków. Zgodnie z wykresami pudełkowymi, są to pojedyncze obserwacje, które znacznie zawyżają statystyki opisowe.

3.3. Analiza rozkładu danych



Rysunek 10. Histogramy dla obu stanów cywilnych

Histogramy obrazują wyraźne przesunięcie dominanty. Płace dla osób w związku małżeńskim skupiają się powyżej 1000 dolarów, a u osób stanu wolnego poniżej tej wartości. Co więcej, bardzo dobrze widać dlaczego odchylenie standardowe w drugiej grupie jest mniejsze. Praktycznie nie występują obserwacje powyżej 2200 dolarów, a maksymalna wartość wynosi w przybliżeniu 3000 dolarów.



Rysunek 11. Porównanie gęstości i dystrybuant dla obu stanów cywilnych

Zbliżone kształty wykresów sugerują, że zarobki dla obu stanów cywilnych mają ten sam rozkład, ale z różnymi parametrami.

4. Podsumowanie

Poznane metody statystyczne pozwoliły na dogłębną analizę danych rzeczywistych. Odpowiedni podział i badanie obserwacji pod wieloma względami odpowiedziały na pytanie jak poszczególne czynniki wpływają na pensję pracownika. Powyższa praca również pokazała jak ważne jest kompleksowe podejście do tematu. Wizualizacja danych oraz takie procesy jak winsoryzacja lub średnia ucinana mają ogromne znaczenie w procesie wyciągania wniosków.