P O L I T E C H N I K A R Z E S Z O W S K A
im. Ignacego Łukasiewicza
WYDZIAŁ MATEMATYKI i FIZYKI STOSOWANEJ

Zarobki Analityków Danych

Adrianna Rapa 173204

Statystyczna Analiza Danych - projekt
kierunek studiów: Inżynieria i Analiza Danych

Rzeszów 2024

Spis treści

1 Opis użytych danych

Projekt realizowany w ramach przedmiotu „Statystyczna Analiza Danych” na kierunku inżynieria i analiza danych, semestr czwarty, grupa L5. Algorytm jest zapisany w środowisku R w programie RStudio.

1.1 Źródło danych:

Dane zostały pozyskane z platformy Kaggle, (https://www.kaggle.com/datasets/zain280/data-science-salaries/data).

1.2 Charakterystyka danych:

Zestaw danych obejmuje informacje o wynagrodzeniach w obszarze analityki danych, zapewnia wgląd w trendy wynagrodzeń w dziedzinie analityki danych w różnych branżach, lokalizacjach, poziomach doświadczenia i stanowiskach. Zawiera 11 kolumn, 607 wierszy, każda obserwacja reprezentuje wynagrodzenie jednego pracownika na określonym stanowisku. Zawiera różnorodne dane dotyczące: stanowisk, lokalizacji, typów zatrudnienia, wielkości firm itp.

1. Rok pracy: Rok lub zakres lat, dla których zebrane są dane dotyczące wynagrodzeń w dziedzinie nauki o danych.
2. Poziom doświadczenia: Informacja o poziomie doświadczenia wymaganym lub posiadanych przez osoby na każdej roli. Może to obejmować kategorie takie jak poziom podstawowy, średnio zaawansowany, zaawansowany lub doświadczony w danej dziedzinie.
3. Rodzaj zatrudnienia: Opisuje, czy pracownik jest zatrudniony na pełny etat, część etatu, pracuje jako freelancer, kontrakt itp.
4. Stanowisko: Etykiety opisowe wskazujące konkretne stanowisko lub rolę w dziedzinie nauki o danych, np. analityk danych, inżynier uczenia maszynowego itp.
5. Wynagrodzenie: Wartości liczbowe przedstawiające roczne lub miesięczne wynagrodzenie na każdym stanowisku.
6. Waluta wynagrodzenia: Jednostka walutowa, w jakiej podane są wartości wynagrodzenia.
7. Wynagrodzenie w USD: Wartość wynagrodzenia przeliczona na dolary amerykańskie.
8. Miejsce zamieszkania pracownika: Lokalizacja, w której pracownik mieszka.
9. Czy pracuje zdalnie: Informacja, czy pracownik pracuje zdalnie, czy też z lokalizacji firmy.
10. Lokalizacja firmy: Lokalizacja geograficzna, w której znajduje się firma, dla której pracuje dana osoba.
11. Wielkość firmy: Informacje o wielkości firmy zatrudniającej, często podzielone na kategorie według liczby pracowników lub przychodów.

	RokPracy	PoziomDoświadczenia	RodzajZatrudnienia	Stanowisko	Wynagrodzenie	WalutaWynagrodzenia	Wynagrodzenie_w_USD	MiejsceZamieszkaniaPracownika	PracaZdalna	MiejsceFirmy	RozmiarFirmy
1	2020	Średnio zaawansowany	Pełny etat	Analityk Danych	70000	EUR	79833	Niemcy	0	Niemcy	Duża
2	2020	Zaawansowany	Pełny etat	Specjalista Uczenia Maszynowego	260000	USD	260000	Japonia	0	Japonia	Mała
3	2020	Zaawansowany	Pełny etat	Inżynier Big Data	85000	GBP	109024	Wielka Brytania	50	Wielka Brytania	Średnia
4	2020	Średnio zaawansowany	Pełny etat	Analityk Danych Produktowych	20000	USD	20000	Honduras	0	Honduras	Mała
5	2020	Zaawansowany	Pełny etat	Inżynier Uczenia Maszynowego	150000	USD	150000	Stany Zjednoczone	50	Stany Zjednoczone	Duża
6	2020	Początkujący	Pełny etat	Analityk Danych	72000	USD	72000	Stany Zjednoczone	100	Stany Zjednoczone	Duża
7	2020	Zaawansowany	Pełny etat	Lider Analityków Danych	190000	USD	190000	Stany Zjednoczone	100	Stany Zjednoczone	Mała
8	2020	Średnio zaawansowany	Pełny etat	Analityk Danych	11000000	HUF	35735	Węgry	50	Węgry	Duża
9	2020	Średnio zaawansowany	Pełny etat	Analityk Danych Biznesowych	135000	USD	135000	Stany Zjednoczone	100	Stany Zjednoczone	Duża
10	2020	Zaawansowany	Pełny etat	Lider Inżynieri Danych	125000	USD	125000	Nowa Zelandia	50	Nowa Zelandia	Mała
11	2020	Początkujący	Pełny etat	Analityk Danych	45000	EUR	51321	Francja	0	Francja	Mała
12	2020	Średnio zaawansowany	Pełny etat	Analityk Danych	3000000	INR	40481	Indie	0	Indie	Duża
13	2020	Początkujący	Pełny etat	Analityk Danych	35000	EUR	39916	Francja	0	Francja	Średnia
14	2020	Średnio zaawansowany	Pełny etat	Lider Analizy Danych	87000	USD	87000	Stany Zjednoczone	100	Stany Zjednoczone	Duża
15	2020	Średnio zaawansowany	Pełny etat	Analityk Danych	85000	USD	85000	Stany Zjednoczone	100	Stany Zjednoczone	Duża
16	2020	Średnio zaawansowany	Pełny etat	Analityk Danych	8000	USD	8000	Pakistan	50	Pakistan	Duża
17	2020	Początkujący	Pełny etat	Inżynier Danych	4450000	JPY	41689	Japonia	100	Japonia	Mała
18	2020	Zaawansowany	Pełny etat	Inżynier Big Data	100000	EUR	114047	Polska	100	Wielka Brytania	Mała
19	2020	Początkujący	Pełny etat	Konsultant Analizy Danych	423000	INR	5707	Indie	50	Indie	Średnia

Rysunek 1 - Kolumny znajdujące się w bazie

1.3 Uzasadnienie wyboru danych:

Wybór tych danych jest uzasadniony ich istotnością dla analizy trendów rynkowych dotyczących wynagrodzeń w branży analityki danych. Dzięki tym danym można lepiej zrozumieć strukturę płacową, czynniki wpływające na wysokość wynagrodzenia oraz różnice regionalne i sektorowe.

2 Wprowadzenie danych

Wprowadzenie danych rozpoczęto od wczytania bazy danych z pliku CSV za pomocą funkcji read.csv. Plik zawiera informacje o wynagrodzeniach pracowników związanych z dziedziną nauki o danych. Po wczytaniu danych, usuwana jest pierwsza kolumna, która zawiera niepotrzebne indeksy lub identyfikatory. Następnie następuje zmiana nagłówków kolumn na bardziej intuicyjne i zrozumiałe, co ułatwia dalszą analizę danych.

Kolejnym krokiem jest dekodowanie kodów państw na angielskie nazwy dla kolumn "MiejsceZamieszkaniaPracownika" i "MiejsceFirmy". Następnie angielskie nazwy stanowisk są zamieniane na polskie tłumaczenia**.**

W dalszej kolejności, angielskie skróty oznaczające poziom doświadczenia, rozmiar firmy i rodzaj zatrudnienia są zamieniane na polskie odpowiedniki. To kolejne przekształcenie, które zwiększa zrozumienie danych, szczególnie dla osób nieobeznanych z terminologią angielską.

Po przeprowadzeniu tych operacji, następuje utworzenie ramki danych zawierającej wybrane kolumny, co ogranicza się do tych najistotniejszych dla analizy. Następnie usuwane są odstające wartości wynagrodzeń, które mogą zakłócać analizę danych lub modelowanie. Dzięki temu uzyskujemy bardziej spójne i wiarygodne dane do analizy.

W bazie danych występują wartości wynagrodzeń w kolumnie "Wynagrodzenie", które mogą być wyrażone zarówno rocznie, jak i miesięcznie. Jednakże sposób przedstawienia tych wartości nie jest jednoznaczny, co może wprowadzać pewne zamieszanie**.** Ostatnim krokiem jest podzielenie wysokich wartości wynagrodzeń rocznych przez 12 i zaokrąglenie ich do najbliższej całkowitej liczby, aby uzyskać wartości miesięczne. To konieczne działanie, które ułatwia porównywanie wynagrodzeń na różnych poziomach czasowych.

W skrócie, wprowadzenie danych obejmuje proces wczytania, oczyszczenia i przekształcenia danych w celu uzyskania spójnej, zrozumiałej i wiarygodnej bazy danych, gotowej do dalszej analizy i modelowania.

	PoziomDoświadczenia	Stanowisko	Wynagrodzenie_w_USD	MiejsceFirmy
1	Średnio zaawansowany	Analityk Danych	79833	Niemcy
2	Zaawansowany	Specjalista Uczenia Maszynowego	260000	Japonia
3	Zaawansowany	Inżynier Big Data	109024	Wielka Brytania
4	Średnio zaawansowany	Analityk Danych Produktowych	20000	Honduras
5	Zaawansowany	Inżynier Uczenia Maszynowego	150000	Stany Zjednoczone
6	Początkujący	Analityk Danych	72000	Stany Zjednoczone
7	Zaawansowany	Lider Analityków Danych	190000	Stany Zjednoczone
8	Średnio zaawansowany	Analityk Danych	35735	Węgry
9	Średnio zaawansowany	Analityk Danych Biznesowych	135000	Stany Zjednoczone
10	Zaawansowany	Lider Inżynieri Danych	125000	Nowa Zelandia
11	Początkujący	Analityk Danych	51321	Francja
12	Średnio zaawansowany	Analityk Danych	40481	Indie
13	Początkujący	Analityk Danych	39916	Francja
14	Średnio zaawansowany	Lider Analizy Danych	87000	Stany Zjednoczone
15	Średnio zaawansowany	Analityk Danych	85000	Stany Zjednoczone
16	Średnio zaawansowany	Analityk Danych	8000	Pakistan
17	Początkujący	Inżynier Danych	41689	Japonia
18	Zaawansowany	Inżynier Big Data	114047	Wielka Brytania
19	Początkujący	Konsultant Analizy Danych	5707	Indie
20	Średnio zaawansowany	Lider Inżynieri Danych	56000	Stany Zjednoczone
21	Średnio zaawansowany	Inżynier Uczenia Maszynowego	43331	Chiny
22	Średnio zaawansowany	Analityk Danych Produktowych	6072	Indie
23	Zaawansowany	Inżynier Danych	47899	Grecja
24	Średnio zaawansowany	Analityk biznesowy	98000	Stany Zjednoczone
25	Średnio zaawansowany	Lider Analityków Danych	115000	Zjednoczone Emiraty Arabskie
26	Doświadczony	Dyrektor Analizy Danych	325000	Stany Zjednoczone
27	Początkujący	Badacz Naukowy	42000	Holandia
28	Zaawansowany	Inżynier Danych	33511	Meksyk
29	Początkujący	Analityk Danych Biznesowych	100000	Stany Zjednoczone

Rysunek 2 - Wybrane kolumny

3 Wyznaczone parametry opisowe

Została przeprowadzana analiza różnych statystyk dla wybranych zarobków, takich jak średnia, odchylenie standardowe, współczynnik zmienności, mediana, maksimum, minimum, rozstęp, wariancja, kwantyle, dominanta, wskaźnik asymetrii.

3.1 Średnia

10389,61 USD

Średnia arytmetyczna wartości wynagrodzeń w całym zbiorze danych. Jest to suma wszystkich wynagrodzeń podzielona przez liczbę obserwacji. Średnia dostarcza ogólnej informacji na temat przeciętnego wynagrodzenia w badanym zbiorze.

```
> #Średnia
> srednia <- mean(wynagrodzenia_AD$Wynagrodzenie_w_USD)
> cat("Średnia: ", srednia, "\n")
Średnia:  10389.61
```

3.2 Odchylenie standardowe

5355,607 USD

Mierzy stopień rozproszenia wartości wynagrodzeń wokół średniej. Im większe odchylenie standardowe, tym większa zmienność wynagrodzeń. Jest to miara rozrzutu wartości wokół średniej.

```
> #Odchylenie standardowe
> odchylenie_std <- sd(wynagrodzenia_AD$Wynagrodzenie_w_USD)
> cat("Odchylenie standardowe: ", odchylenie_std, "\n")
Odchylenie standardowe:  5355.607
```

3.3 Mediana

9408 USD

Wartość środkowa w uporządkowanym ciągu danych. Dzieli zbiór na dwie równe części: 50% obserwacji znajduje się powyżej, a 50% poniżej tej wartości. Mediana jest mniej wrażliwa na wartości skrajne niż średnia.

```
> #Mediana
> mediana <- median(wynagrodzenia_AD$Wynagrodzenie_w_USD)
> cat("Mediana: ", mediana, "\n")
Mediana:  9408
```

3.4 Dominanta

8333 USD

Wartość, która pojawia się najczęściej w zbiorze danych.

```
> #Dominanta
> dominanta <- names(sort(table(wynagrodzenia_AD$Wynagrodzenie_w_USD), decreasing = TRUE))[1]
> cat("Dominanta: ", dominanta, "\n")
Dominanta:  8333
```

3.5 Rozstęp

27215 USD

Różnica między maksymalną a minimalną wartością wynagrodzenia w USD w ramce danych wynagrodzenia_AD.

```
> #Rozstęp
> rozstep <- max(wynagrodzenia_AD$Wynagrodzenie_w_USD) - min(wynagrodzenia_AD$Wynagrodzenie_w_USD)
> cat("Rozstęp: ", rozstep, "\n")
Rozstęp:  27215
```

3.6 Kwartyle

- **I - 2356 USD**
- **II - 6282 USD**
- **III - 9408 USD**
- **IV - 13333 USD**
- **V - 29751 USD**

Reprezentują punkty podziału danych na cztery równoliczne części. Pierwszy kwartyl (Q1) to 25% najniższych wynagrodzeń, mediana (Q2) to 50% wynagrodzeń, a trzeci kwartyl (Q3) to 75% wynagrodzeń. Kwartyle dostarczają informacji o rozkładzie danych.

```
> #Kwartyle
> kwartyle <- quantile(wynagrodzenia_AD$Wynagrodzenie_w_USD)
> cat("Kwartyle: ", kwartyle, "\n")
Kwartyle:  2536 6282 9408 13333 29751
```

3.7 Minimum

2536 USD

Najmniejsza wartość wynagrodzenia w badanym zbiorze danych**.**

```
> #Minimum
> minimum <- min(wynagrodzenia_AD$Wynagrodzenie_w_USD)
> cat("Minimum: ", minimum, "\n")
Minimum:  2536
```

3.8 Maksimum

29751 USD

Największa wartość wynagrodzenia w badanym zbiorze danych.

```
> #Maksimum
> maksimum <- max(wynagrodzenia_AD$Wynagrodzenie_w_USD)
> cat("Maksimum: ", maksimum, "\n")
Maksimum:  29751
```

3.9 Wariancja

28682523

Mierzy zmienność danych poprzez średnią kwadratów różnic między wartościami a ich średnią. Większa wariancja oznacza większą zmienność danych.

```
> #Wariancja
> wariancja <- var(wynagrodzenia_AD$Wynagrodzenie)
> cat("Wariancja: ", wariancja, "\n")
Wariancja:  28682523
```

3.10 Współczynnik zmienności

51,5477%

Określa stopień zmienności wynagrodzeń w stosunku do ich średniej wartości. Jest to stosunek odchylenia standardowego do średniej, wyrażony jako procent. Ta miara pozwala porównywać zmienność wynagrodzeń w różnych zbiorach danych.

```
> #Współczynnik zmienności
> wspolczynnik_zmiennosci <- odchylenie_std / srednia * 100
> cat("Współczynnik zmienności: ", wspolczynnik_zmiennosci, "%\n")
Współczynnik zmienności:  51.5477 %
```

3.11 **Momenty centralne:**

- Pierwszy moment centralny: Średnia arytmetyczna różnic między wartościami wynagrodzeń a ich średnią.
- Drugi moment centralny (odchylenie standardowe): Średnia arytmetyczna kwadratów różnic między wartościami wynagrodzeń a ich średnią.
- Trzeci moment centralny: Średnia arytmetyczna sześcianów różnic między wartościami wynagrodzeń a ich średnią.
- Czwarty moment centralny (kurtoza): Średnia arytmetyczna czwartych potęg różnic między wartościami wynagrodzeń a ich średnią.

```
> #Momenty centralne
> #Pierwszy moment centralny (powinien być bliski zera)
> moment_centralny_1 <- sum((wynagrodzenia_AD$Wynagrodzenie - srednia)^1) / length(wynagrodzenia_AD$Wynagrodzenie)
> cat("Pierwszy moment centralny: ", moment_centralny_1, "\n")
Pierwszy moment centralny:  -1.518355e-14
> #Drugi moment centralny (to odchylenie standardowe)
> moment_centralny_2 <- sum((wynagrodzenia_AD$Wynagrodzenie - srednia)^2) / length(wynagrodzenia_AD$Wynagrodzenie)
> cat("Drugi moment centralny (odchylenie standardowe): ", moment_centralny_2, "\n")
Drugi moment centralny (odchylenie standardowe):  28634639
> #Trzeci moment centralny
> moment_centralny_3 <- sum((wynagrodzenia_AD$Wynagrodzenie - srednia)^3) / length(wynagrodzenia_AD$Wynagrodzenie)
> cat("Trzeci moment centralny: ", moment_centralny_3, "\n")
Trzeci moment centralny:  146153814445
> #Czwarty moment centralny (to kurtoza)
> moment_centralny_4 <- sum((wynagrodzenia_AD$Wynagrodzenie - srednia)^4) / length(wynagrodzenia_AD$Wynagrodzenie)
> cat("Czwarty moment centralny (kurtoza): ", moment_centralny_4, "\n")
Czwarty moment centralny (kurtoza):  3.084586e+15
```

3.12 **Współczynnik asymetrii**

9514459

Miara asymetrii rozkładu danych. Dodatnia wartość wskazuje na przewagę wartości większych od średniej, a ujemna na przewagę wartości mniejszych.

```
> #Współczynnik asymetrii
> wspolczynnik_asymetrii <- sum((wynagrodzenia_AD$Wynagrodzenie - srednia)^3) / (length(wynagrodzenia_AD$Wynagrodzenie) * odchylenie_std^3)
> cat("Współczynnik asymetrii: ", wspolczynnik_asymetrii, "\n")
Współczynnik asymetrii:  0.9514459
```

3.13 **Wskaźnik asymetrii**

rozkład prawoskośny

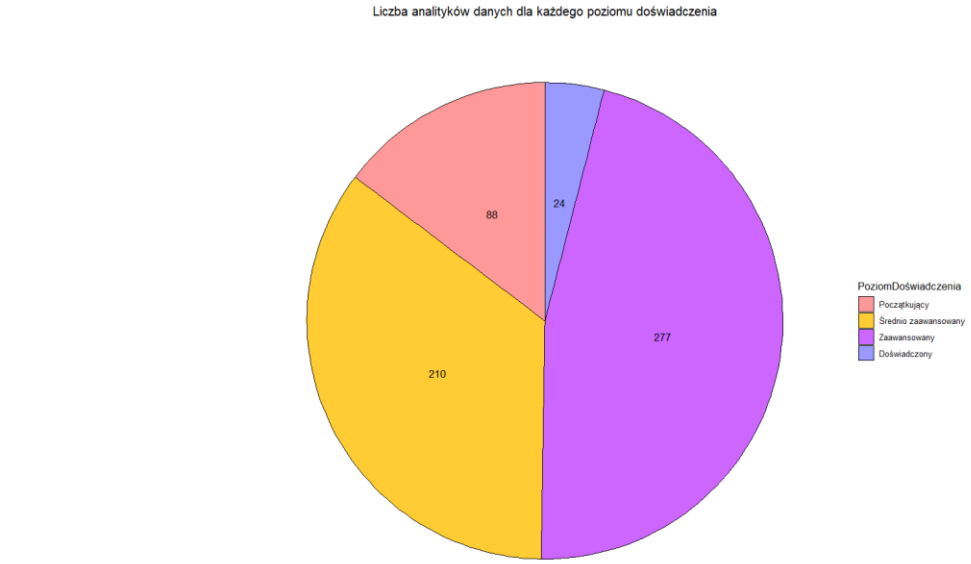
Wartość wskazująca na kierunek asymetrii rozkładu danych: prawoskośny, lewoskośny lub symetryczny.

```
> #Wskaźnik asymetrii z wykorzystaniem współczynnika asymetrii
> if (wspolczynnik_asymetrii > 0) {
+   cat("Rozkład prawoskośny (skośność w prawo).\n")
+ } else if (wspolczynnik_asymetrii < 0) {
+   cat("Rozkład lewoskośny (skośność w lewo).\n")
+ } else {
+   cat("Rozkład symetryczny.\n")
+ }
Rozkład prawoskośny (skośność w prawo).
```

4 Graficzna prezentacja danych

4.1 Wykres kołowy

```r
#Graficzna prezentacja danych
# Zainstalowanie i załadowanie pakietów
install.packages("ggplot2")
library(ggplot2)

### WYKRES KOŁOWY
#Liczba analityków danych dla każdego poziomu doświadczenia
doswiadczenie_counts <- table(wynagrodzenia_AD$PoziomDoświadczenia)

#Przekształcenie do data frame
doswiadczenie_df <- as.data.frame(doswiadczenie_counts)
colnames(doswiadczenie_df) <- c("PoziomDoświadczenia", "LiczbaAnalitykow")

#Paleta kolorów
paleta_kolorow <- c("#FF9999", "#ffcc33", "#cc66ff", "#9999ff")

#Wykres kołowy
wykres_kolowy <- ggplot(doswiadczenie_df, aes(x = "", y = LiczbaAnalitykow, fill = PoziomDoświadczenia)) +
  geom_bar(stat = "identity", width = 1, color = "black", size = 0.5) +  # Dodanie czarnych obramowań
  coord_polar(theta = "y") +
  labs(title = "Liczba analityków danych dla każdego poziomu doświadczenia", x = NULL, y = NULL) +
  theme_minimal() +
  theme(axis.text.x = element_blank(), axis.ticks = element_blank(), panel.grid = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  scale_fill_manual(values = paleta_kolorow) +  # Ustawienie kolorów zdefiniowanych wcześniej
  geom_text(aes(label = LiczbaAnalitykow), position = position_stack(vjust = 0.5))

#Wyświetlenie wykresu
print(wykres_kolowy)
```

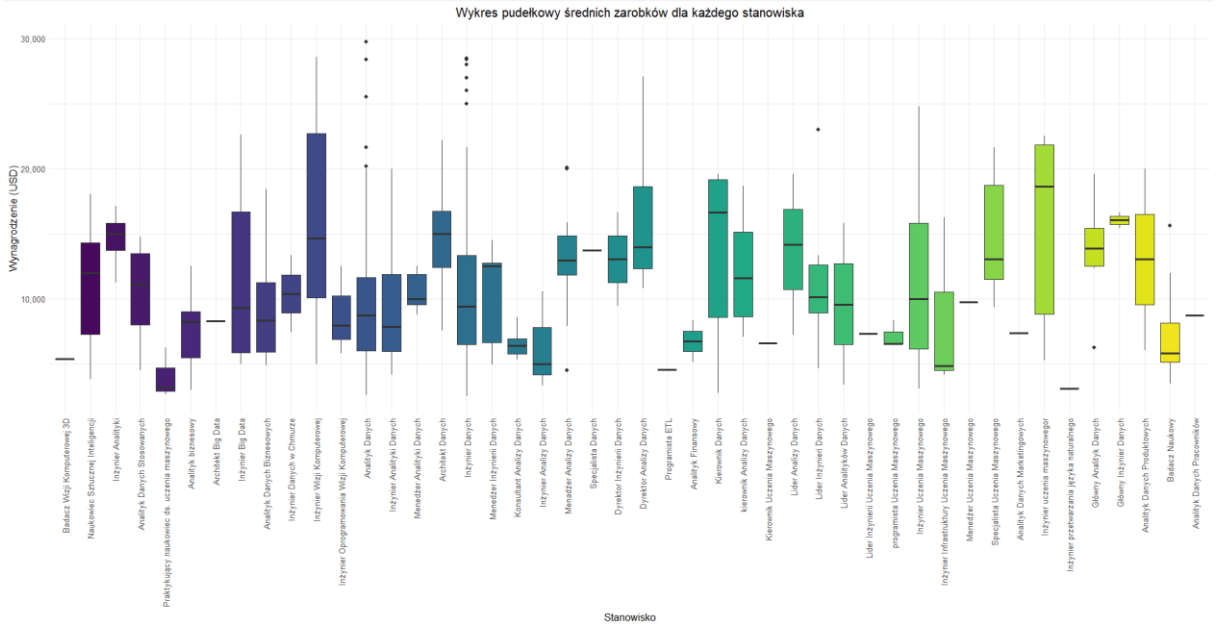


Rysunek 3 - Wykres kołowy

Wykres kołowy przedstawia liczbę analityków danych dla różnych poziomów doświadczenia. Każdy segment koła reprezentuje jeden z poziomów doświadczenia, a jego wielkość odpowiada liczbie analityków danych na danym poziomie. Kolory segmentów są używane do odróżnienia poszczególnych poziomów doświadczenia. Dodatkowo, wartości numeryczne, czyli liczby analityków danych, są umieszczone na wykresie za pomocą etykiet tekstowych. Dzięki temu wykresowi można łatwo porównać liczbę analityków danych na różnych poziomach doświadczenia.

4.2 Wykres pudełkowy

```
####WYKRES PUDEŁKOWY ŚREDNICH ZAROBKOW DLA KAŻDEGO STANOWISKA
wykres_pudełkowy <- ggplot(wynagrodzenia_AD, aes(x = Stanowisko, y = Wynagrodzenie_w_USD, fill = Stanowisko)) +
  geom_boxplot() +
  labs(title = "Wykres pudełkowy średnich zarobków dla każdego stanowiska", x = "Stanowisko", y = "Wynagrodzenie (USD)") +
  scale_y_continuous(labels = scales::comma) +  # Formatowanie osi y bez notacji wykładniczej
  scale_fill_viridis_d() +  # Ustawienie automatycznie generowanych kolorów
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        legend.position = "none")  # Usunięcie legendy

#Wyświetlenie wykresu
print(wykres_pudełkowy)
```



Rysunek 4 - Wykres pudełkowy

Wykres przedstawia rozkład średnich zarobków dla różnych stanowisk w danych dotyczących wynagrodzeń. Każde pudełko na wykresie pudełkowym reprezentuje rozkład danych dla danego stanowiska.

Oś x: Reprezentuje różne stanowiska.

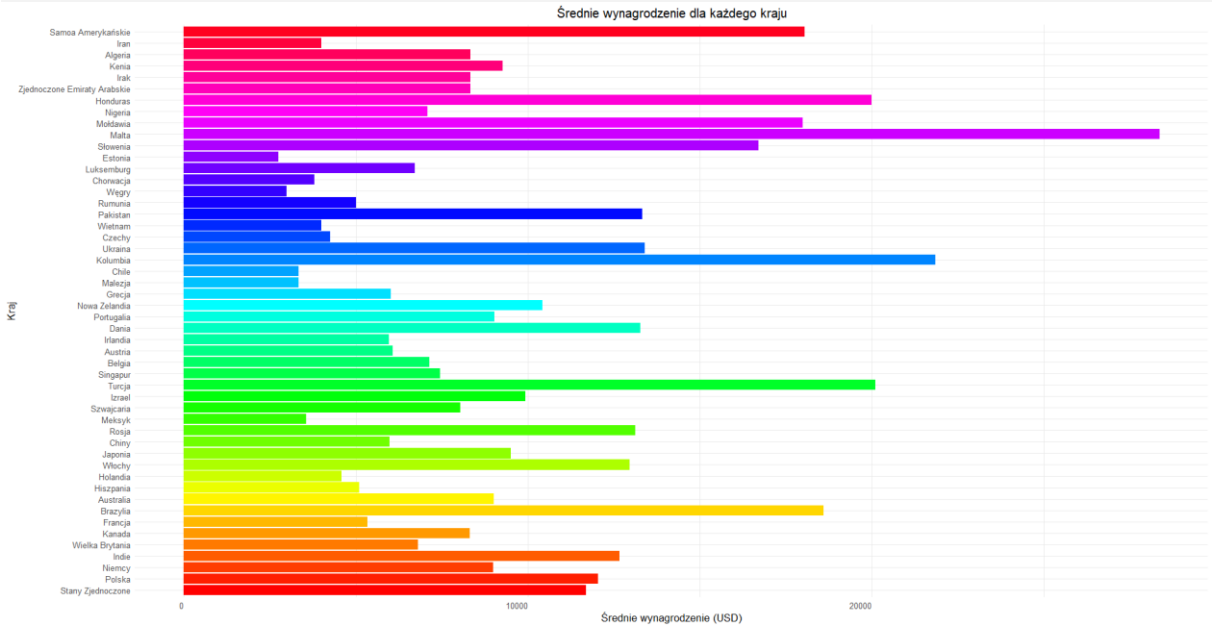
Oś y: Reprezentuje zarobki w dolarach USD.

Pudełka (Boxplot): Każde pudełko przedstawia kwartyle danych: dolny kwartyl (Q1), mediana (linia środkowa pudełka) oraz górny kwartyl (Q3). Linie "wąsów" (whiskers) rozciągają się z każdego pudełka i pokazują zakres danych poza kwartylami. Punkty lub punkty odstające mogą być również wyświetlane, reprezentując wartości odstające spoza zakresu wskazanego przez wąsy.

Wykres pudełkowy jest przydatny do porównywania rozkładów danych między różnymi kategoriami, w tym przypadku różnymi stanowiskami. Pomaga zidentyfikować zróżnicowanie w zarobkach między różnymi stanowiskami, w tym medianę, zakres oraz występowanie wartości odstających.

4.3 Wykres słupkowy poziomy

```
###WYKRES SŁUPKOWY ŚREDNICH ZAROBKÓW W KAŻDYM KRAJU
#Średnie wynagrodzenie dla każdego kraju
srednie_wynagrodzenie_kraju <- aggregate(Wynagrodzenie_w_USD ~ MiejsceFirmy, data = wynagrodzenia_AD, FUN = mean)

#Sortowanie według średniego wynagrodzenia malejąco
srednie_wynagrodzenie_kraju <- srednie_wynagrodzenie_kraju[order(-srednie_wynagrodzenie_kraju$Wynagrodzenie_w_USD), ]


#Paleta kolorów dla każdego kraju
kolor_kraju <- rainbow(length(unique(srednie_wynagrodzenie_kraju$MiejsceFirmy)))

#Wykres słupkowy ze średnim wynagrodzeniem dla każdego kraju
wykres_slupkowy <- ggplot(srednie_wynagrodzenie_kraju, aes(x = MiejsceFirmy, y = Wynagrodzenie_w_USD, fill = MiejsceFirmy)) +
  geom_bar(stat = "identity") +
  labs(title = "Średnie wynagrodzenie dla każdego kraju", x = "Kraj", y = "Średnie wynagrodzenie (USD)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1),  # Pochylenie etykiet osi x
        plot.title = element_text(hjust = 0.5),  # Wyśrodkowanie tytułu
        legend.position = "none") +  # Usunięcie legendy
  scale_fill_manual(values = kolor_kraju) +  # Ustawienie różnych kolorów dla każdego kraju
  coord_flip()  # Obrócenie wykresu

#Wyświetlenie wykresu
print(wykres_slupkowy)
```



Rysunek 5 - Wykres słupkowy poziomy

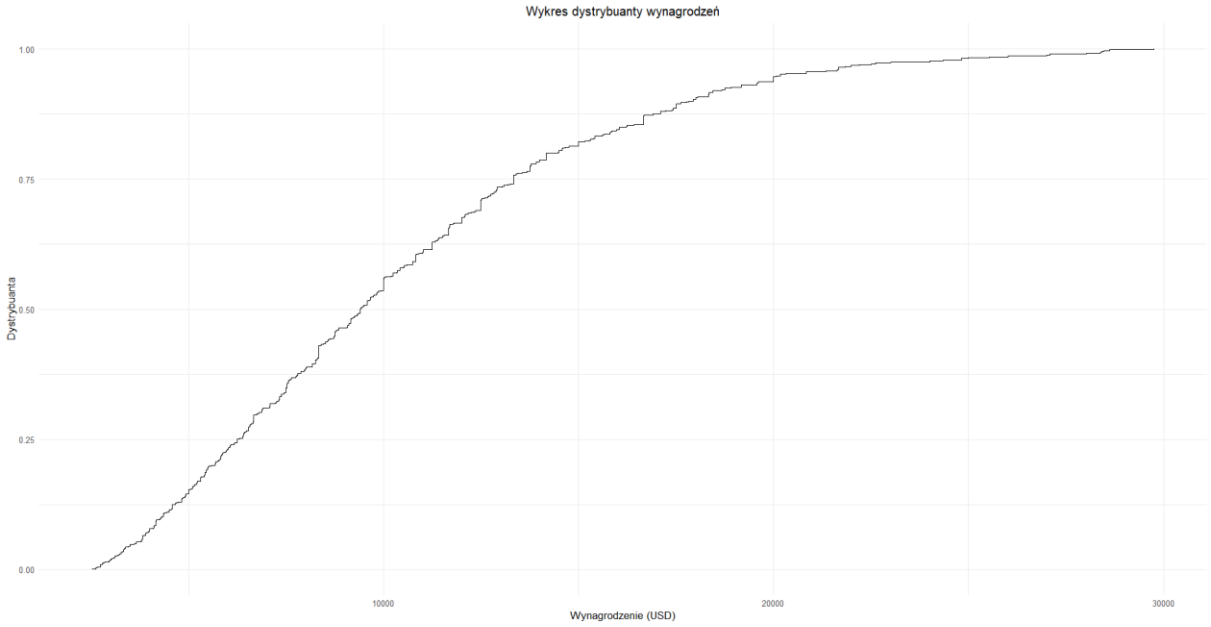
Wykres słupkowy przedstawia średnie wynagrodzenie dla każdego kraju, bazując na danych zawartych w analizowanej bazie danych. Oś pozioma (x) reprezentuje poszczególne kraje, natomiast oś pionowa (y) przedstawia średnie wynagrodzenie wyrażone w dolarach amerykańskich (USD).

Każdy słupek na wykresie reprezentuje średnie wynagrodzenie dla danego kraju. Służą one do porównania poziomów wynagrodzeń między różnymi krajami. Im wyższy słupek, tym wyższe średnie wynagrodzenie w danym kraju.

Kolory słupków odpowiadają poszczególnym krajom, co ułatwia identyfikację, który kraj jest reprezentowany przez dany słupek. Paleta kolorów została wygenerowana automatycznie, aby każdy kraj miał inny kolor, co ułatwia wizualne porównanie danych.

4.4 Wykres dystrybuanty

```
###WYKRES DYSTRYBUANTY
#Wygenerowanie wykresu dystrybuanty
wykres_dystrybuanty <- ggplot(wynagrodzenia_AD, aes(x = Wynagrodzenie_w_USD)) +
  stat_ecdf(geom = "step", pad = FALSE) +  # Dodanie wykresu dystrybuanty
  labs(title = "Wykres dystrybuanty wynagrodzeń", x = "Wynagrodzenie (USD)", y = "Dystrybuanta") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))  # Wyśrodkowanie tytułu

#Wyświetlenie wykresu
print(wykres_dystrybuanty)
```



Rysunek 6 - Wykres dystrybuanty

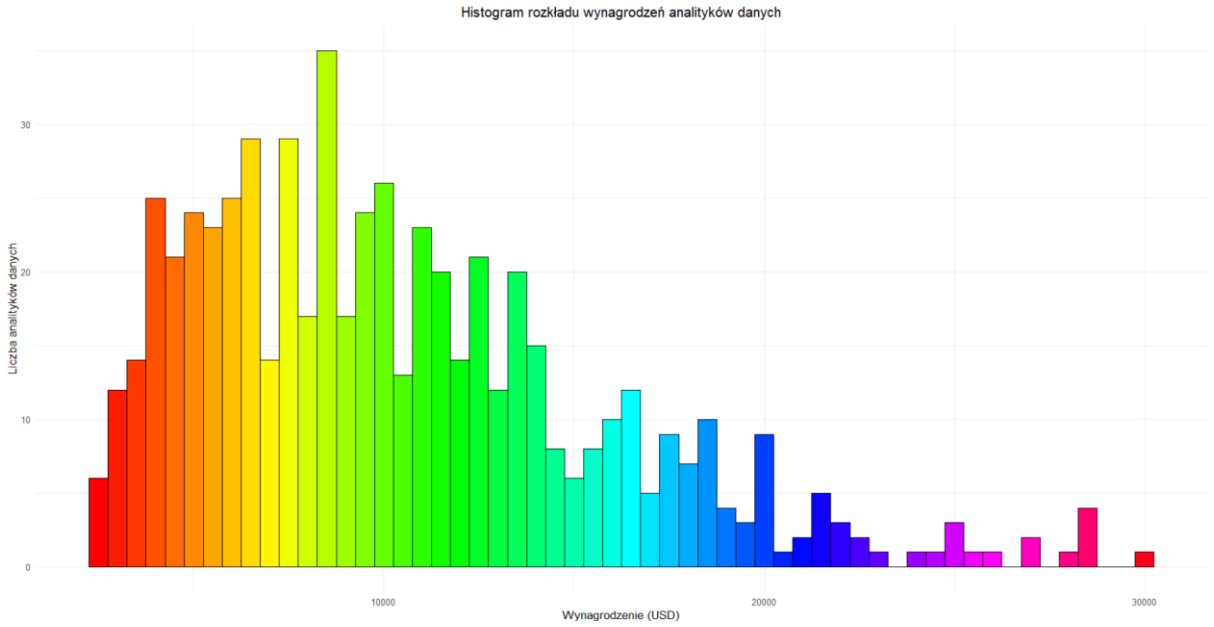
Wykres przedstawia dystrybuantę empiryczną dla wynagrodzeń zawartych w analizowanej bazie danych. Oś pozioma (x) reprezentuje wartości wynagrodzeń wyrażone w dolarach amerykańskich (USD), natomiast oś pionowa (y) przedstawia wartości dystrybuanty, które są interpretowane jako prawdopodobieństwo, że losowo wybrane wynagrodzenie będzie mniejsze lub równe danej wartości na osi x.

Wykres został stworzony za pomocą funkcji stat_ecdf() z argumentem geom = "step", co oznacza, że przedstawia on dystrybuantę w formie schodkowej, gdzie wartości dystrybuanty zmieniają się skokowo na podstawie kolejnych wartości wynagrodzeń.

Wykres ten pozwala na szybkie zrozumienie rozkładu wartości wynagrodzeń w analizowanej próbie danych. Im bardziej wartości dystrybuanty zbliżają się do wartości 1, tym większy odsetek wynagrodzeń znajduje się poniżej danej wartości na osi x.

4.5 Histogram

```
###HISTOGRAM
#Wygenerowanie palety kolorów tęczy
n_colors <- 56 # Określenie liczby kolorów w palecie
rainbow_palette <- colorRampPalette(colors = rainbow(n_colors))

#Histogram rozkładu wynagrodzeń analityków danych dla wszystkich krajów naraz
wykres_histogram <- ggplot(wynagrodzenia_AD, aes(x = Wynagrodzenie_w_USD, fill = factor(..x..))) +
  geom_histogram(binwidth = 500, color = "black") +
  scale_fill_manual(values = rainbow_palette(n_colors)) +
  labs(title = "Histogram rozkładu wynagrodzeń analityków danych", x = "Wynagrodzenie (USD)", y = "Liczba analityków danych") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")  # Usunięcie legendy i wyśrodkowanie tytułu

#Wyświetlenie wykresu
print(wykres_histogram)
```



Rysunek 7 - Histogram

Jest to histogram rozkładu wynagrodzeń analityków danych, gdzie osią poziomą (x) są wartości wynagrodzeń wyrażone w dolarach amerykańskich (USD), a osią pionową (y) jest liczba analityków danych.

Każdy słup na wykresie reprezentuje przedział wartości wynagrodzeń, a jego wysokość odpowiada liczbie analityków danych, których wynagrodzenie znajduje się w danym przedziale. Szerokość słupków jest równa szerokości przedziałów, co oznacza, że są one równoodległe od siebie.

Aby zapewnić czytelność wykresu, wartości wynagrodzeń zostały podzielone na przedziały o stałej szerokości, co umożliwia łatwe porównanie rozkładu wynagrodzeń między różnymi grupami.

5 Hipotezy statystyczne

5.1 Hipoteza 1

Opis: Istnieje statystycznie istotna różnica między średnimi wynagrodzeniami analityków danych w Polsce i USA.

Hipoteza zerowa (H0):

Średnie wynagrodzenie analityków danych w Polsce jest takie samo jak średnie wynagrodzenie analityków danych w Stanach Zjednoczonych. Matematycznie:

$H0 : \mu\text{Polska} = \mu\text{USA}$

Średnie wynagrodzenie analityków danych w Polsce jest takie samo jak w Stanach Zjednoczonych.

Hipoteza alternatywna(H1):

Średnie wynagrodzenie analityków danych w Polsce różni się od średniego wynagrodzenia analityków danych w Stanach Zjednoczonych. Matematycznie:

$H1 : \mu\text{Polska} \neq \mu\text{USA}$

Średnie wynagrodzenie analityków danych w Polsce różni się od wynagrodzenia w Stanach Zjednoczonych.

Test:

```
> #Hipotezy
> 'H1: Istnieje statystycznie istotna różnica między średnimi wynagrodzeniami analityków danych w Polsce i USA.'
[1] "H1: Istnieje statystycznie istotna różnica między średnimi wynagrodzeniami analityków danych w Polsce i USA."
> # Filtracja danych dla Polski i Stanów Zjednoczonych
> dane_polska <- subset(wynagrodzenia_AD, MiejsceFirmy == "Polska")
> dane_usa <- subset(wynagrodzenia_AD, MiejsceFirmy == "Stany Zjednoczone")
> # Wyciągnięcie wynagrodzeń
> wynagrodzenia_polska <- dane_polska$Wynagrodzenie_w_USD
> wynagrodzenia_usa <- dane_usa$Wynagrodzenie_w_USD
> # Test normalności Shapiro-Wilka
> shapiro_polska <- shapiro.test(wynagrodzenia_polska)
> shapiro_usa <- shapiro.test(wynagrodzenia_usa)
> library("car")
> levene_test <- leveneTest(Wynagrodzenie_w_USD ~ MiejsceFirmy, data = wynagrodzenia_AD)
> # Test t-studenta
> test_t <- t.test(wynagrodzenia_polska, wynagrodzenia_usa, var.equal = TRUE)
> # Alternatywnie, test t-studenta z korektą Welch'a
> test_t_welch <- t.test(wynagrodzenia_polska, wynagrodzenia_usa, var.equal = FALSE)
> # Wyniki testu t-studenta
> print(test_t)

        Two Sample t-test

data:  wynagrodzenia_polska and wynagrodzenia_usa
t = 0.14818, df = 349, p-value = 0.8823
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4106.501  4775.684
sample estimates:
mean of x mean of y
 12032.75  11698.16

> print(test_t_welch)

        Welch Two Sample t-test

data:  wynagrodzenia_polska and wynagrodzenia_usa
t = 0.056539, df = 3.0095, p-value = 0.9585
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -18465.47  19134.65
sample estimates:
mean of x mean of y
 12032.75  11698.16
```

Interpretacja:

Wartość t: W obu testach wartość t jest bardzo niska (0.14818 dla testu standardowego i 0.056539 dla testu z korektą Welch'a). Oznacza to, że różnica między średnimi wynagrodzeniami w Polsce i USA jest minimalna.

Wartość p-value:

W teście standardowym wartość p-value wynosi 0.8823.

W teście z korektą Welch'a wartość p-value wynosi 0.9585.

W obu przypadkach wartość p-value jest znacznie większa niż standardowy poziom istotności 0.05. Oznacza to, że nie możemy odrzucić hipotezy zerowej. Brak dowodów na istnienie statystycznie istotnej różnicy między średnimi wynagrodzeniami analityków danych w Polsce i USA.

Przedział ufności:

Dla testu standardowego przedział ufności wynosi od -4106.501 do 4775.684 USD.

Dla testu z korektą Welch'a przedział ufności wynosi od -18465.47 do 19134.65 USD.

W obu przypadkach przedział ufności obejmuje zero, co wskazuje na brak statystycznie istotnej różnicy między średnimi wynagrodzeniami w obu krajach.

Średnie wynagrodzenia: Średnie wynagrodzenie analityków danych w Polsce wynosi 12,032.75 USD, a w USA 11,698.16 USD. Różnica między średnimi wynagrodzeniami jest niewielka.

Wniosek: Na podstawie wyników testu t-studenta (zarówno standardowego, jak i z korektą Welch'a) nie możemy odrzucić hipotezy zerowej. Oznacza to, że nie ma statystycznie istotnej różnicy między średnimi wynagrodzeniami analityków danych w Polsce i Stanach Zjednoczonych. Wyniki sugerują, że średnie wynagrodzenia w obu krajach są porównywalne.

5.2 **Hipoteza 2**

Opis: Istnieje statystycznie istotna różnica między średnimi wynagrodzeniami analityków danych w Polsce i Niemczech.

Hipoteza zerowa (H0):

Średnie wynagrodzenie analityków danych w Polsce jest takie samo jak średnie wynagrodzenie analityków danych w Niemczech. Matematycznie:

$H0 : \mu\text{Polska} = \mu\text{Niemcy}$

Hipoteza alternatywna (H1):

Średnie wynagrodzenie analityków danych w Polsce różni się od średniego wynagrodzenia analityków danych w Niemczech. Matematycznie:

$H1: \mu\text{Polska} \neq \mu\text{Niemcy}$

Test:

```
> # H2: Istnieje statystycznie istotna różnica między średnimi wynagrodzeniami analityków danych w Polsce i Niemczech.
> # Filtracja danych dla Polski i Niemiec
> dane_polska <- subset(wynagrodzenia_AD, MiejsceFirmy == "Polska")
> dane_niemcy <- subset(wynagrodzenia_AD, MiejsceFirmy == "Niemcy")
> # Wyciągnięcie wynagrodzeń
> wynagrodzenia_polska <- dane_polska$Wynagrodzenie_w_USD
> wynagrodzenia_niemcy <- dane_niemcy$Wynagrodzenie_w_USD
> # Test normalności Shapiro-Wilka
> shapiro_polska <- shapiro.test(wynagrodzenia_polska)
> shapiro_niemcy <- shapiro.test(wynagrodzenia_niemcy)
> # Test równości wariancji Levene'a
> library("car")
> levene_test <- leveneTest(Wynagrodzenie_w_USD ~ MiejsceFirmy, data = wynagrodzenia_AD)
> # Test t-studenta
> test_t <- t.test(wynagrodzenia_polska, wynagrodzenia_niemcy, var.equal = TRUE)
> # Alternatywnie, test t-studenta z korektą Welch'a
> test_t_welch <- t.test(wynagrodzenia_polska, wynagrodzenia_niemcy, var.equal = FALSE)
> # Wyniki testu t-studenta
> print(test_t)

        Two Sample t-test

data:  wynagrodzenia_polska and wynagrodzenia_niemcy
t = 0.87745, df = 30, p-value = 0.3872
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4032.411 10107.554
sample estimates:
mean of x mean of y
12032.750  8995.179

> print(test_t_welch)

        Welch Two Sample t-test

data:  wynagrodzenia_polska and wynagrodzenia_niemcy
t = 0.50573, df = 3.193, p-value = 0.6459
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15439.95  21515.09
sample estimates:
mean of x mean of y
12032.750  8995.179
```

Interpretacja:

Wyniki testów t-studenta i t-studenta z korektą Welch'a wskazują na brak istotnej różnicy między średnimi wynagrodzeniami analityków danych w Polsce i Niemczech.

Dla testu t-studenta:

Wartość p wynosi 0.3872, co oznacza, że nie ma wystarczających dowodów na odrzucenie hipotezy zerowej o równości średnich wynagrodzeń.

Dodatkowo, przedział ufności dla różnicy średnich wynagrodzeń zawiera wartość zero, co dodatkowo potwierdza brak istotnej różnicy.

Podobnie, dla testu t-studenta z korektą Welch'a:

Wartość p wynosi 0.6459, również nie daje podstaw do odrzucenia hipotezy zerowej.

Przedział ufności dla różnicy średnich wynagrodzeń również zawiera zero.

W obu przypadkach, średnie wynagrodzenia analityków danych w Polsce i Niemczech są podobne, gdyż nie ma istotnej statystycznie różnicy między nimi.

6 Komentarz do uzyskanych wyników

Analiza została przeprowadzona na zbiorze danych dotyczących wynagrodzeń analityków danych. Zostały wykonane różne operacje na danych, takie jak przekształcenie nazw krajów i stanowisk na polskie odpowiedniki, usunięcie odstających wartości, a także przeprowadzenie obliczeń statystycznych i graficzna prezentacja wyników.

- Przekształcenie danych: Dane zostały poddane kilku przekształceniom, m.in. zmianie nazw krajów i stanowisk na polskie tłumaczenia oraz dekodowaniu skrótów na pełne nazwy odpowiadających im poziomów doświadczenia, rozmiarów firm i rodzajów zatrudnienia.

- Usuwanie odstających wartości: Zastosowano metodę usuwania odstających wartości, aby wyeliminować nieprawdopodobnie wysokie lub niskie wynagrodzenia. Wykorzystano kryterium 3-krotnego odchylenia standardowego od średniej.

- Analiza statystyczna: Przeprowadzono analizę opisową, obejmującą obliczenie średniej, odchylenia standardowego, mediany, dominanty, rozstępu, kwartyli, minimum, maksimum, wariancji, współczynnika zmienności, momentów centralnych, współczynnika asymetrii oraz wskaźnika asymetrii.

- Graficzna prezentacja danych: Wyniki zostały zaprezentowane w formie wykresów kołowych, pudełkowych, słupkowych, dystrybuanty oraz histogramu, co pozwala na wizualizację i lepsze zrozumienie rozkładu wynagrodzeń oraz innych cech w zbiorze danych.

- Wnioski: Analiza pozwala na zrozumienie struktury wynagrodzeń analityków danych w różnych kontekstach, takich jak poziom doświadczenia, rodzaj zatrudnienia, lokalizacja geograficzna czy rodzaj firmy. Ponadto, eliminacja odstających wartości oraz obliczenia statystyczne pozwalają na lepsze zrozumienie ogólnego obrazu i charakterystyki danych.

Podsumowując, wykonana analiza dostarcza głębokiej wiedzy na temat wynagrodzeń analityków danych oraz ich zależności od różnych czynników, co może być użyteczne w podejmowaniu decyzji biznesowych i planowaniu kariery zawodowej.

7 Opis użytych funkcji środowiska R

Funkcje:

- read.csv() - Funkcja służąca do wczytywania danych z pliku CSV do środowiska R.
- install.packages() - Funkcja instalująca pakiety R z CRAN (Comprehensive R Archive Network) lub innych źródeł.
- library() - Funkcja służąca do ładowania zainstalowanych pakietów R do bieżącej sesji.
- countrycode() - Funkcja z pakietu "countrycode", która służy do przekształcania kodów państw na ich pełne angielskie nazwy.
- factor() - Funkcja konwertująca wektor na zmienną jakościową (faktor), która jest wykorzystywana do reprezentacji kategorii lub poziomów zmiennych jakościowych.
- colnames() - Funkcja służąca do nadawania nowych nazw kolumnom w ramce danych.
- subset() - Funkcja służąca do wybierania podzbiorów danych na podstawie określonych kryteriów.
- t.test(): Wykonuje test t-studenta dla dwóch prób niezależnych.
- mean() - Funkcja obliczająca średnią arytmetyczną wartości w wektorze lub kolumnie ramki danych.
- shapiro.test(): Wykonuje test normalności Shapiro-Wilka.
- sd() - Funkcja obliczająca odchylenie standardowe wartości w wektorze lub kolumnie ramki danych.
- leveneTest(): Wykonuje test równości wariancji Levene'a.
- aggregate() - Funkcja służąca do obliczania statystyk podsumowujących (np. średniej, sumy) w oparciu o grupowanie danych.
- which() - Funkcja zwracająca indeksy elementów, które spełniają określone warunki.
- geom_bar() - Funkcja z pakietu "ggplot2", która rysuje słupkowy wykres na podstawie danych.
- coord_polar() - Funkcja z pakietu "ggplot2", która zmienia współrzędne na polarne, co pozwala na rysowanie wykresów kołowych.
- labs() - Funkcja z pakietu "ggplot2", która służy do nadawania etykiet wykresom.
- theme() - Funkcja z pakietu "ggplot2", która pozwala na dostosowywanie wyglądu elementów wykresu, takich jak tło, osie czy tytuły.
- scale_fill_manual() - Funkcja z pakietu "ggplot2", która pozwala na manualne ustawienie kolorów wypełnienia na wykresie.
- geom_text() - Funkcja z pakietu "ggplot2", która dodaje tekst do wykresu.
- stat_ecdf() - Funkcja z pakietu "ggplot2", która oblicza empiryczną dystrybuantę kumulatywną danych.
- colorRampPalette() - Funkcja generująca paletę kolorów na podstawie podanej sekwencji kolorów.
- geom_histogram() - Funkcja z pakietu "ggplot2", która rysuje histogram na podstawie danych.
- print - Funkcja służąca do wyświetlania danych lub obiektów w konsoli.
- names -Funkcja zwracająca lub ustawiająca nazwy elementów w wektorze lub liście.
- median - Funkcja obliczająca medianę z wektora liczb.
- max - Funkcja zwracająca największą wartość z wektora liczb.
- min - Funkcja zwracająca najmniejszą wartość z wektora liczb.
- quantile - Funkcja obliczająca kwantyle z wektora liczb.
- sum - Funkcja obliczająca sumę elementów wektora liczb.
- cat - Funkcja służąca do wyświetlania tekstu w konsoli.
- as.data.frame - Funkcja konwertująca obiekt na ramkę danych.
- ggplot - Funkcja do tworzenia wykresów w pakiecie ggplot2.

Biblioteki:

- readxl: Obsługuje import danych z arkuszy kalkulacyjnych MS Excel.
- ggplot2: Obsługuje tworzenie wykresów i wizualizacji danych.
- DescTools: Obsługuje obliczanie statystyk opisowych, takich jak dominanta.
- e1071: Obsługuje obliczanie wskaźników asymetrii i innych statystyk związanych z analizą danych.
- car: Zawiera funkcje statystyczne, w tym test Levene'a.
- stats: Zawiera funkcje statystyczne, takie jak test t-studenta i test Shapiro-Wilka.