Komputerowa Analiza Szeregów Czasowych - raport 1

Stano Szymon 268776, Julia Wołk-Łaniewska 268735

19 lutego 2024

1 Wstęp

W niniejszym raporcie zajmiemy się analizą danych dotyczących wysokości zarobków w zależności od doświadczenia dla 1000 pracowników. Dane zostały zaczerpnięte ze strony: https://www.kaggle.com/datasets/saquib7hussain/experience-salary-dataset/data.

Pojedynczy wiersz zawiera:

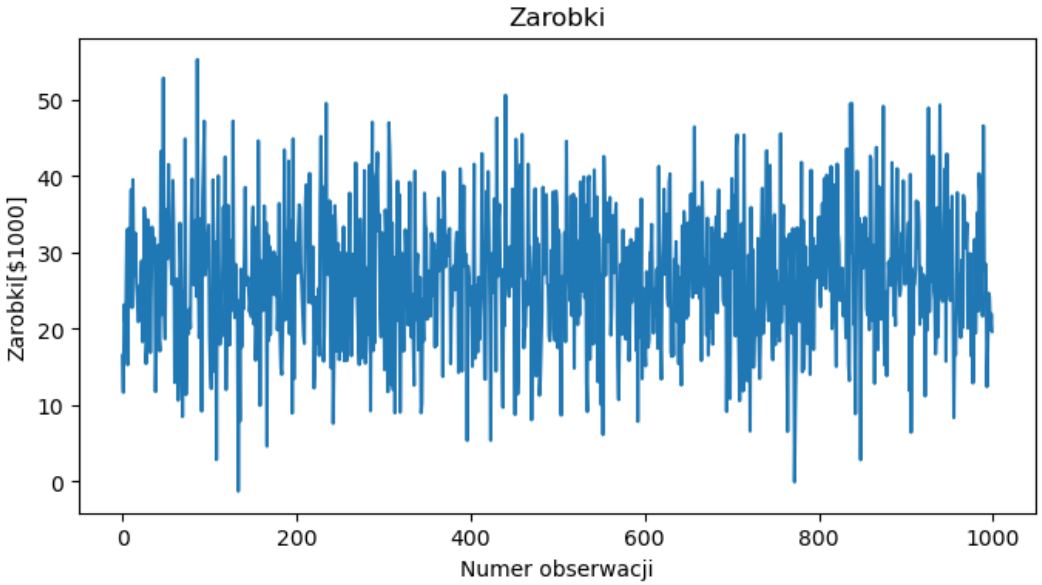
- doświadczenie [miesiąc],
- wynagrodzenie [$1000].

Poniżej zajmiemy się analizą wysokości wynagrodzenia i doświadczenia, jako ciągów zmiennych niezależnych, analizą ich zależności liniowej, predykcją kolejnych danych oraz analizą residuów. Wszystkie obliczenia oraz wykresy zostały wykonane w *Pythonie*.

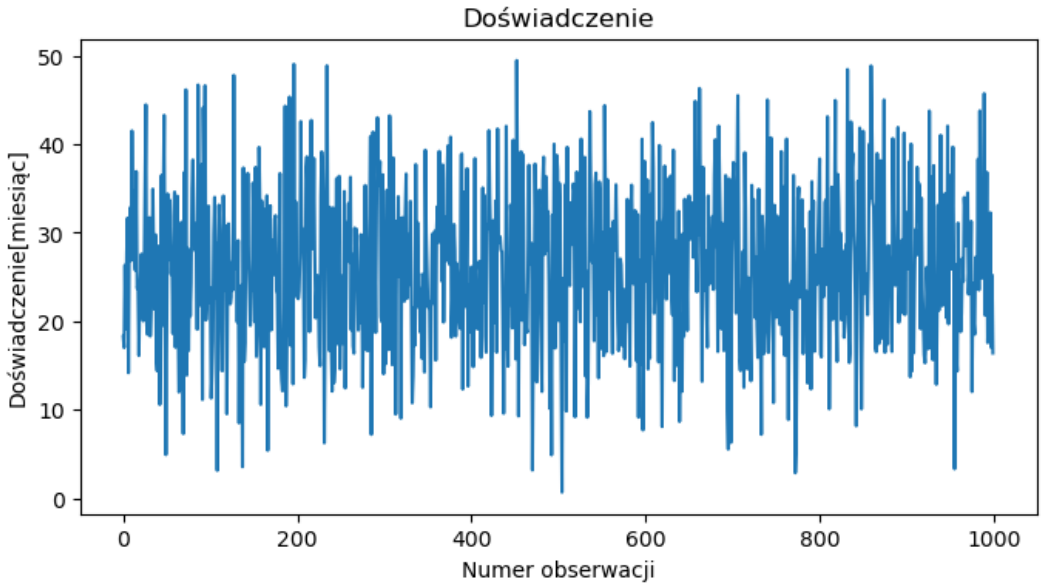
2 Analiza jednowymiarowa

2.1 Filtracja danych

Analizę jednowymiarową danych rozpoczeliśmy od wykonania ich wykresów względem numeru obserwacji (Rysunki 1 i 2). Zauważyliśmy, dzięki temu, że w przypadku wykresu zarobków pojawiają się wartości ujemne. Niemożliwym jest, aby w rzeczywistości wartości zarobków były ujemne, zatem postanowiliśmy odfiltrować te dane i odpowiadające im wartości długości doświadczenia. W ten sposób otrzymaliśmy 998 danych, które będziemy analizować w dalszej części raportu.



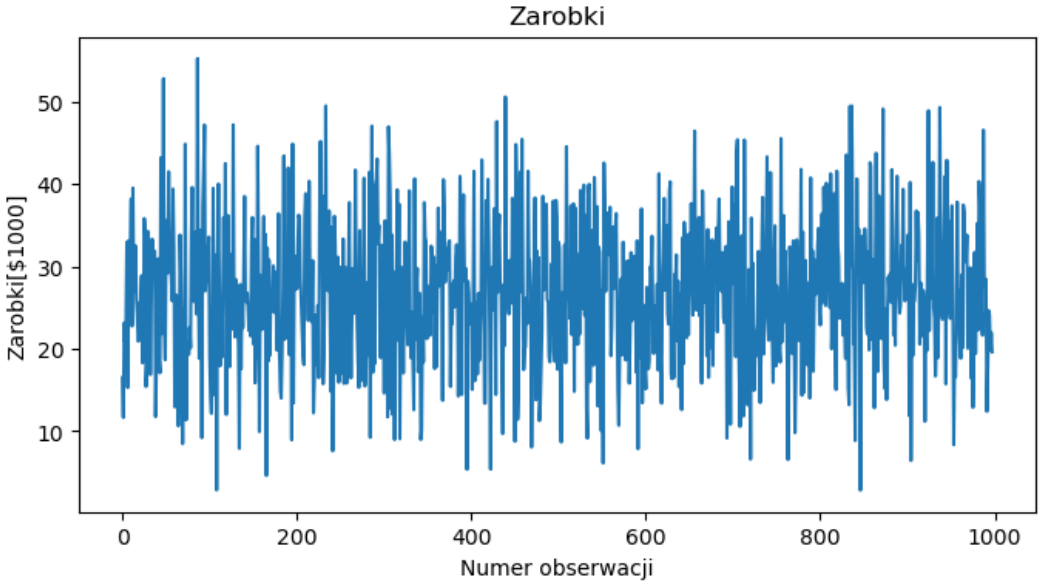
Rysunek 1: Wykres wynagrodzenia



Rysunek 2: Wykres doświadczenia

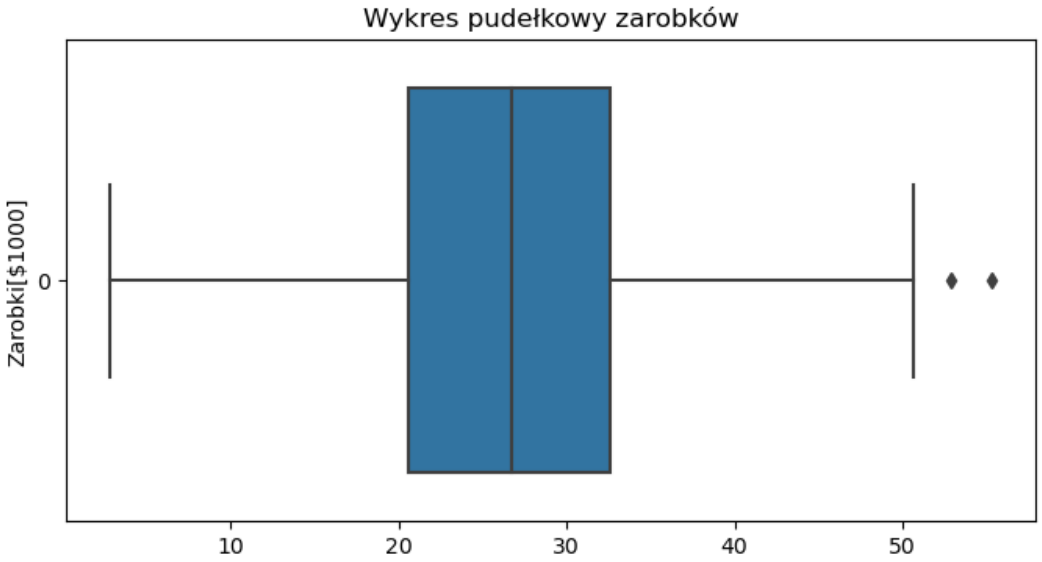
2.2 Wynagrodzenie

W tej części zajmiemy się analizą przefiltrowanych wysokości zarobków. Poniżej umieściliśmy wykres wysokości zarobków w zależności od numeru obseracji (Rysunek 3). Jak możemy zauważyć dane oscylują wokół wartości 25 i odchylają się o około 20. Widzimy również, że najwięcej danych występuje właśnie w okolicy wartości 25 i im bardziej odchylamy się od niej, tym wartości jest coraz mniej. Na pierwszy rzut oka, ze względu na taki wygląd danych, możemy przypuszczać, że mogą one mieć rozkład normalny.

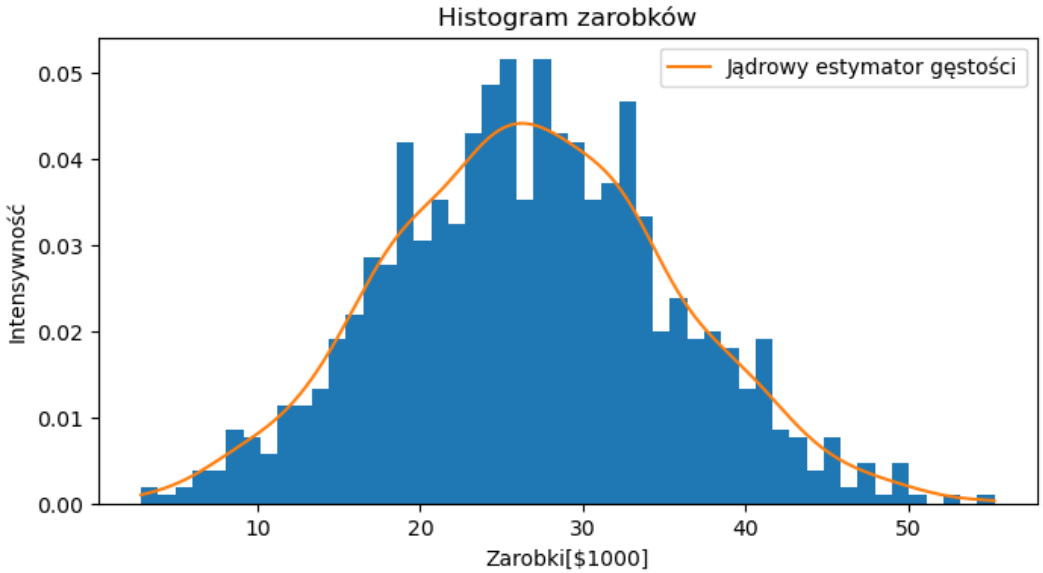


Rysunek 3: Wykres zarobków przefiltrowanych danych

Dalej analizując wykres pudełkowy (Rysunek 4) i histogram (Rysunek 5) możemy dojść do podobnych wniosków. Najwięcej danych jest skumulowanych wokół wartości średniej i rozstęp międzykwartylowy jest stosunkowo niewielki. Widzimy również, że odstających danych jest niewiele.



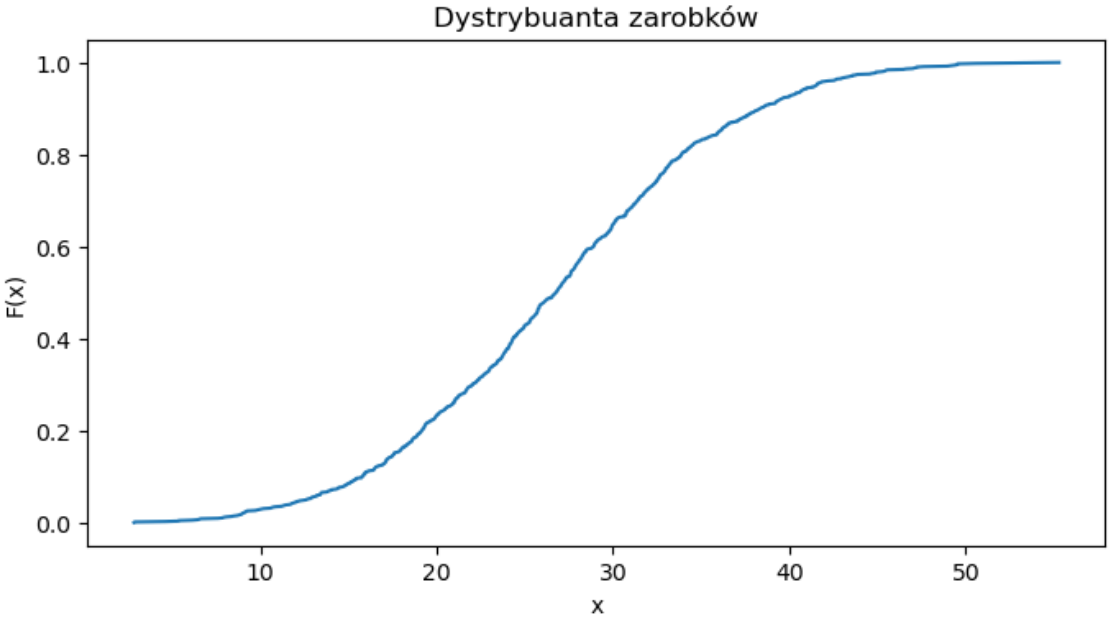
Rysunek 4: Wykres pudełkowy wynagrodzenia



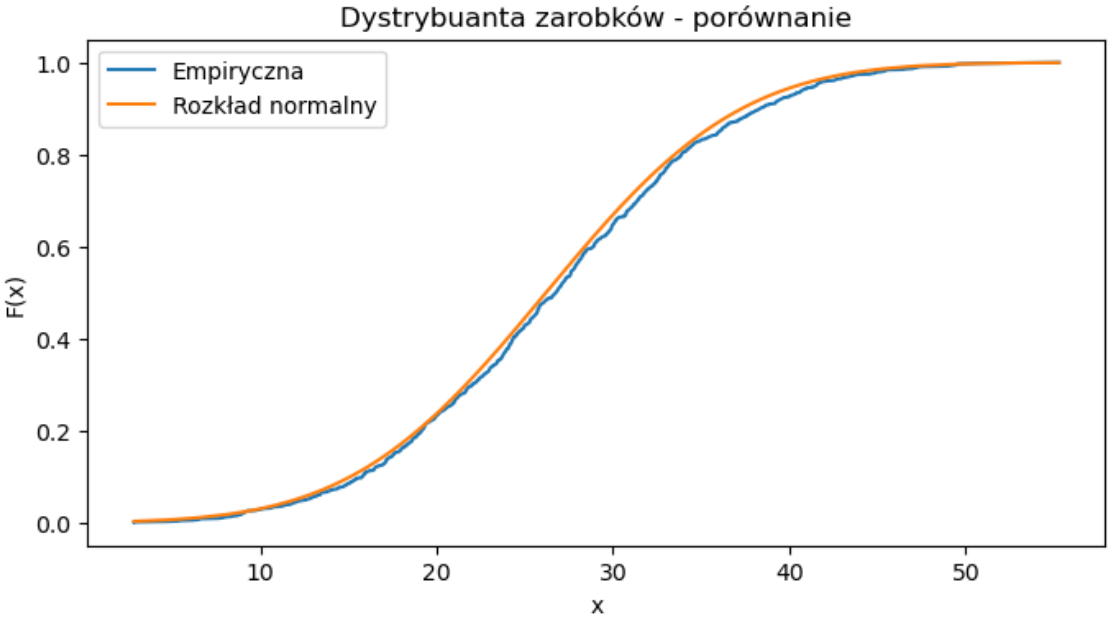
Rysunek 5: Histogram wynagrodzenia

Przyglądając się z kolei dystrybuancie (Rysunek 6) widzimy, że również rośnie ona najszybciej na przedziale [25, 35], a co za tym idzie danych w tym przedziale jest najwięcej i im bardziej oddalamy się od tego obszaru dystrybuanta się wypłaszcza, a więc danych o takich wartościach jest coraz mniej.

Mając przed sobą wykresy dystrybuanty (Rysunek 7) i gęstości (Rysunek 8) danych, widzimy, że wysoce prawdopodobnym jest, że będą one pochodziły z rozkładu normalnego.



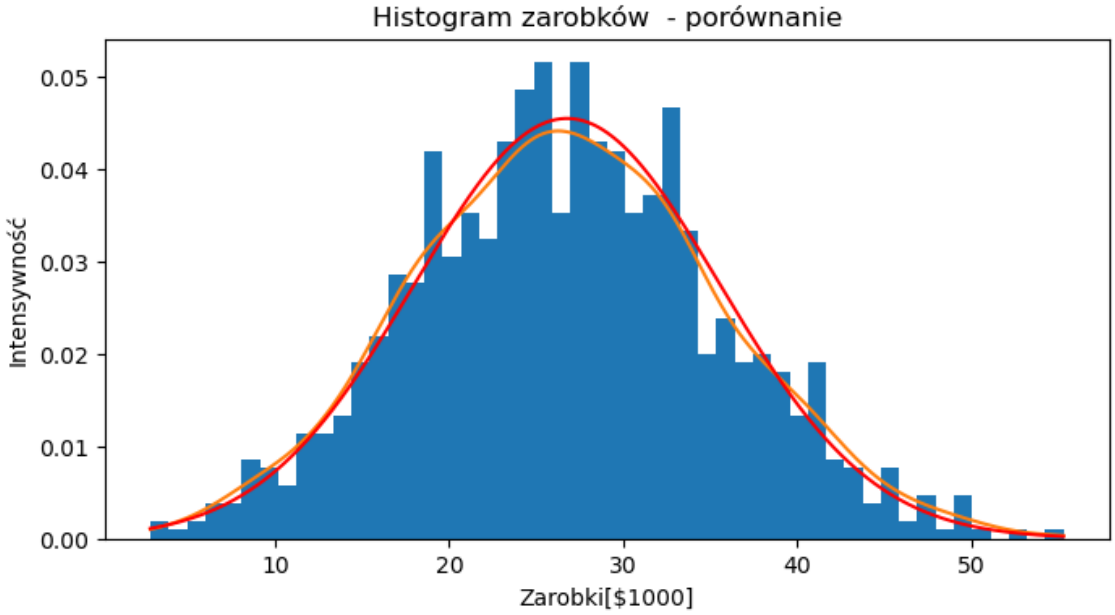
Rysunek 6: Dystrybuanta wynagrodzenia



Rysunek 7: Dystrybuanta wynagrodzenia - porównanie

Dalej na podstawie danych obliczyliśmy ich podstawowe statystyki. Ich wartości umieściliśmy w poniższej tabeli (Tabela 1).

Wyniki potwierdzają nasze dotychczasowe rozważania. Możemy zauważyć, że rozstęp międzykwartylowy jest raczej niewielki w porównaniu z rozstępem, dane skupione są wokół wartości 25 i rozproszone w niewielkim stopniu. Możemy tak wnioskować z wartości wariancji i współczynnika zmienności. Również kurtoza, której wartość jest bliska zeru, informuje nas, że dane nie mają tendencji do dużych odchyleń. Jednak z drugiej strony jej wartośc jest daleka od 3, która cechuje rozkład normalny, co wprowadza w wątpliwość nasze wcześniejsze rozważania. Z kolei współczynnik skośności jest dodatni, z czego możemy wnioskować, że rozkład jest prawostronnie skośny. Jednak jednocześnie jego wartość jest bliska zeru, co może sugerować, że mimo lekkiej skośności rozkład jest dość symetryczny.



Rysunek 8: Histogram zarobków - porównanie

Charakterystyka	Wartość [$1000]
Wartość minimalna	2.832976
Wartość maksymalna	55.297016
Dolny kwartyl (Q1)	20.580347
Mediana (Q2)	26.767413
Górny kwartyl (Q3)	32.619291
Rozstęp międzykwartylowy	12.03894432463703
Rozstęp	52.46404
Wariancja	76.80012507592834
Odchylenie standardowe	8.763568056215934
Kurtoza (dla r.norm. = 3)	-0.1512337065668854
Współczynnik skośności	0.10143819316052366
Współczynnik zmienności	32.727353233031344%

Tabela 1: Charakterystyki liczbowe dla wynagrodzenia.

Następnie obliczyliśmy wartości średnich: arytmetycznej, geometrycznej i harmonicznej (Tabela 2). Ich wyniki umieściliśmy poniżej:

Średnia	Wartość [$1000]
Arytmetyczna	26.777503
Geometryczna	25.122760766466897
Harmoniczna	22.998495217697677

Tabela 2: Średnie dla wynagrodzenia.

Jak możemy zauważyć, średnie oscylują w okolicy wartości mediany, co może oznaczać stosunkowo niedużą ilość danych odstających.

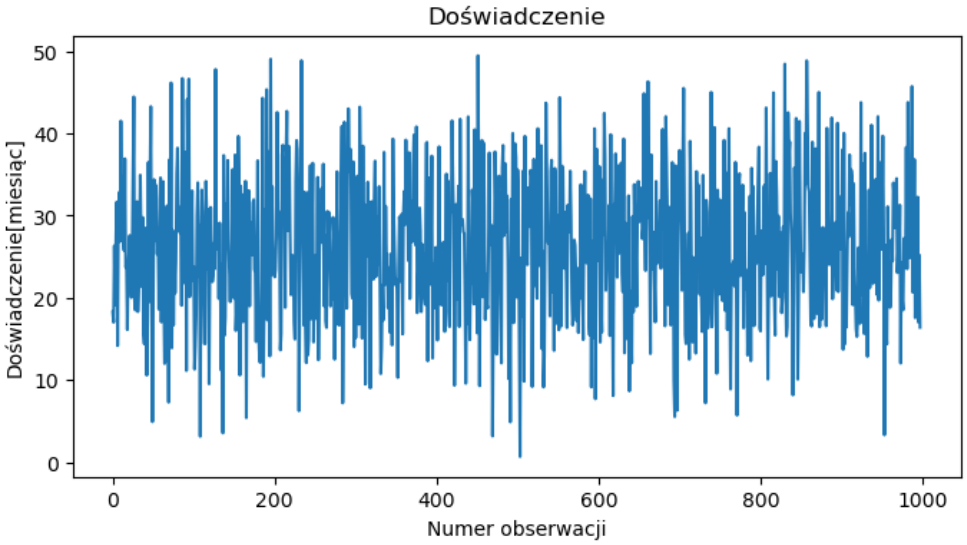
Na koniec przeprowadziliśmy testy statystyczne na normalność analizowanych przez nas danych (Tabela 3):

2.3 Doświadczenie

W tej części dokonaliśmy analogicznej jednowymiarowej analizy dla długości doświadczenia. Poniżej umieściliśmy wykres wysokości zarobków w zależności od numeru obserwacji (Rysunek 9).

Test	Statystyka	P-value
Shapiro-Wilka	0.9983501434326172	0.4625833332538605
Jarque-Bera	2.662600050182806	0.2641336576191245

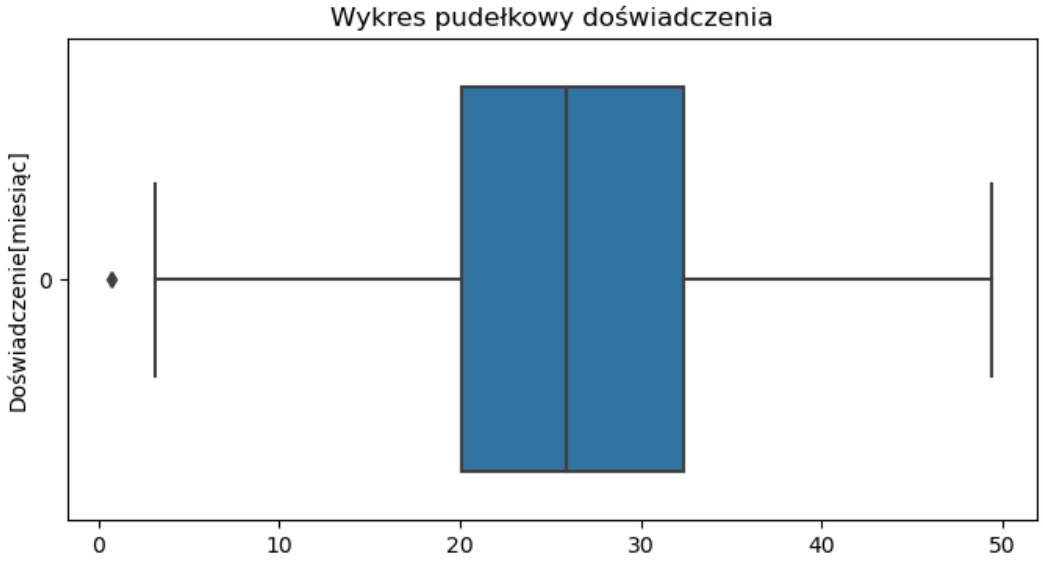
Tabela 3: Caption



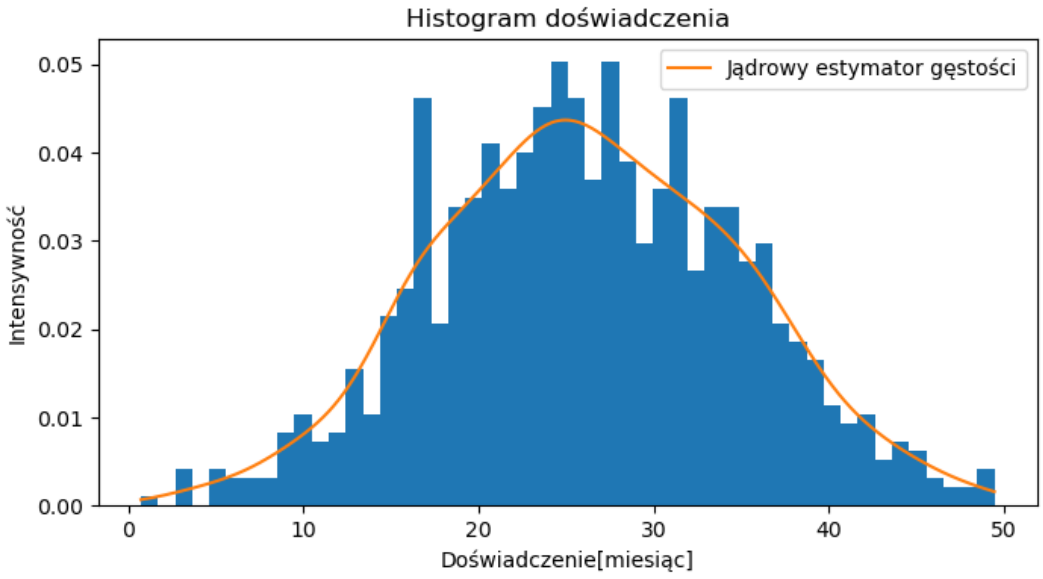
Rysunek 9: Wykres doświadczenia przefiltrowanych danych

Jak widzimy dane wyglądają bardzo podobnie. Skupiają się wokół wartości 30, im bliżej tej wartości, tym zagęszczenie danych jest większe. Z kolei im bardziej się oddalamy, tym wartości, które odstają jest coraz mniej. Nie widzimy też żadnych tendencji względem numeru obserwacji, dane rozkładają się równomiernie na całej długości.

Analizując teraz wykres pudełkowy i histogram umieszczone poniżej (Rysunek 10 i 11) znajdujemy potwierdzenie naszych wcześniejszych rozważań. Widzimy, że dane skupiają się na przedziale [20, 30] i im dalej od tego przedziału tym danych jest mniej. Jednocześnie są one rozproszne w małym stopniu (rozstęp międzykwartylowy niewielki), a wartości odstających jest stosunkowo mało.

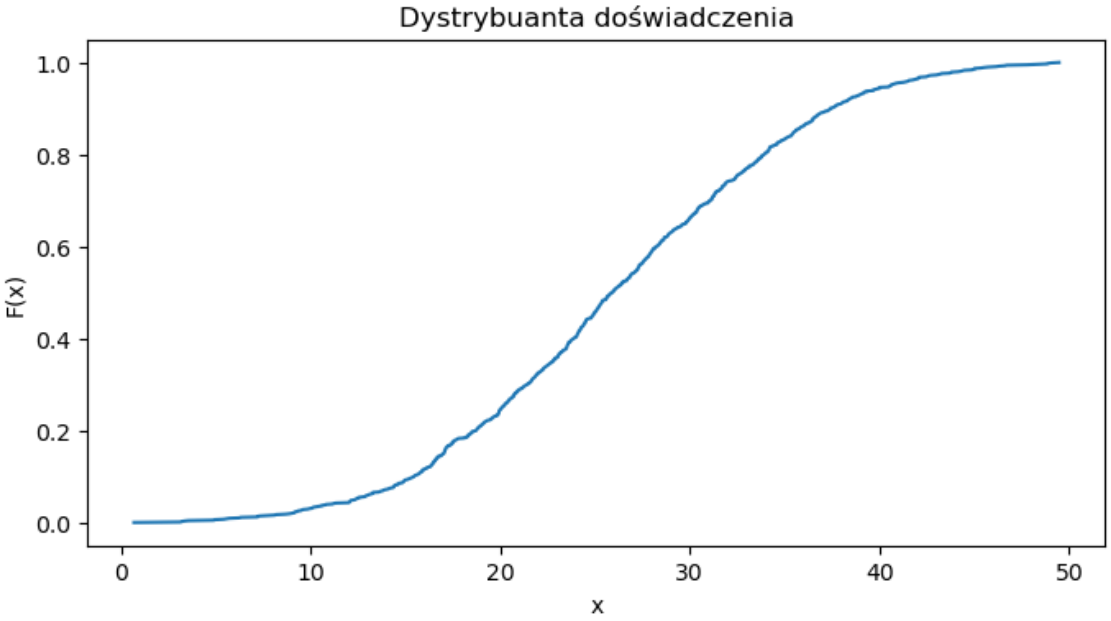


Rysunek 10: Wykres pudełkowy doświadczenia



Rysunek 11: Histogram doświadczenia

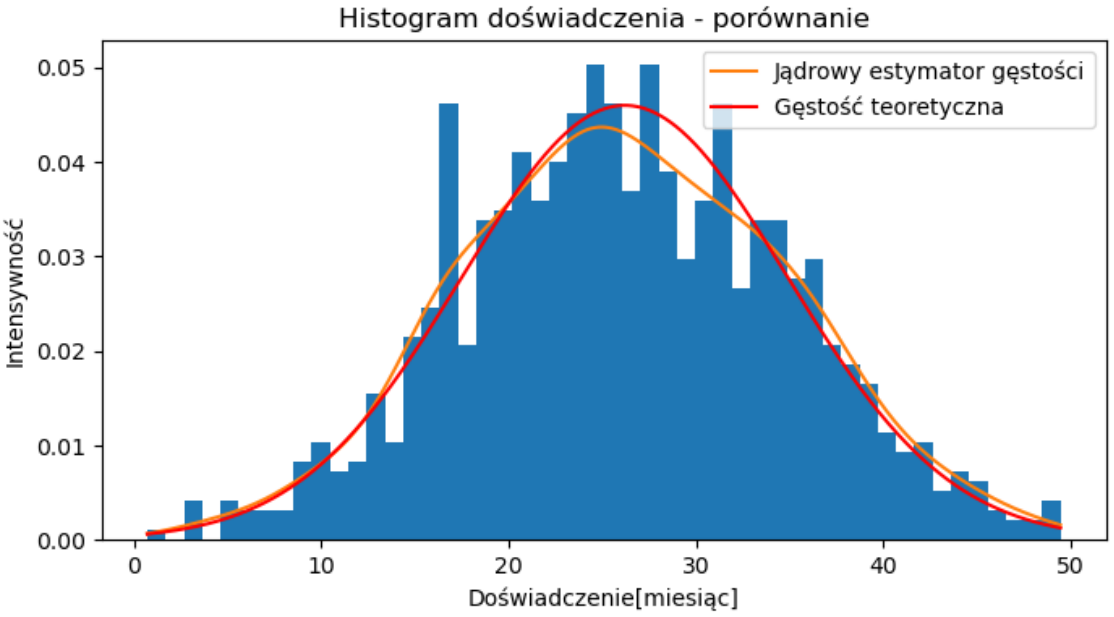
Dalej sporządziliśmy wykres dystrybuanty empirycznej(Rysunek 12).



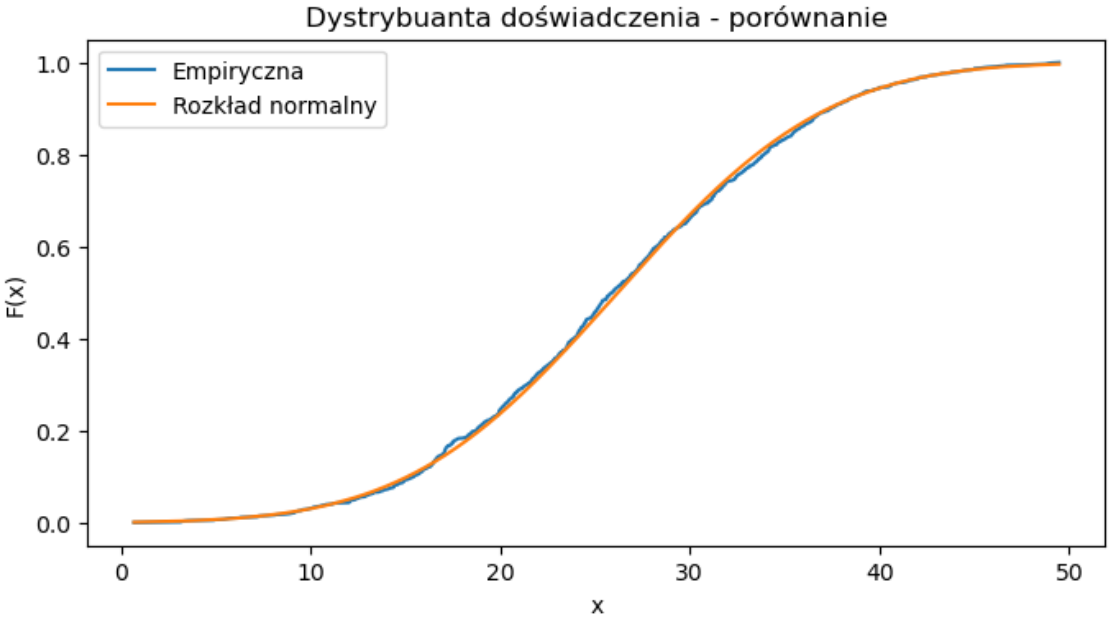
Rysunek 12: Dystrybuanta doświadczenia

Przyglądając się wykresowi dystrybuanty widzimy analogiczne zachowanie. Dystrybuanta rośnie najszybciej na przedziale [20, 30], co oznacza, że jest tam najwięcej danych, im dalej od niego tym funkcja coraz bardziej się wypłaszcza, zatem danych jest coraz mniej.

Analogicznie jak w przypadku danych dotyczących zarobków widzimy, że wartości doświadczenia wykazują zachowanie wskazujące na to, że mają rozkład normalny. Poniżej (Rysunki 13 i 14) umieściliśmy porównanie gęstości i dystrybuanty empirycznych porównanych z funkcjami teoretycznymi rozkładu normalnego.



Rysunek 13: Histogram doświadczenia - porównanie



Rysunek 14: Dystrybuanta doświadczenia - porównanie

Widzimy, że funkcje empiryczne i teoretyczne są bardzo zbliżone, co pozwala przypuszczać, że dane mają rozkład normalny. Dodatkowo sprawdziliśmy to za pomocą testów statystycznych. Wyniki umieściliśmy poniżej (Tabela 4).

Test	Statystyka	P-value
Shapiro-Wilka	0.9978876709938049	0.2401675432920456
Jarque-Bera	3.312489416267184	0.19085434983736668

Tabela 4: Wyniki testów statystycznych

Następnie na podstawie danych obliczyliśmy ich podstawowe statystyki. Ich wartości umieściliśmy w poniższej tabeli (Tabela 5).

Charakterystyka	Wartość [miesiąc]
Wartość minimalna	0.697594
Wartość maksymalna	49.463222
Dolny kwartyl (Q1)	20.090774
Mediana (Q2)	25.891328
Górny kwartyl (Q3)	32.413938
Rozstęp międzykwartylowy	12.323164276029015
Rozstęp	48.765628
Wariancja	75.20759520185975
Odchylenie standardowe	8.672231270086133
Kurtoza (dla r.norm. = 3)	-0.2785240884284437
Współczynnik skośności	0.022822118020436283
Współczynnik zmienności	33.10161063623319%

Tabela 5: Charakterystyki liczbowe dla doświadczenia.

Wartości statystyk potwierdzają tendencje, które zauważyliśmy już wcześniej. Widzimy, że rozstęp międzykwartylowy wynosi około 12 i jest niewielki w porównaniu z rozstępem. Dolny kwartyl wynosi około 20, a górny 32, co pokrywa się z przedziałem największego zagęszczenia danych, który obserowaliśmy na wcześniejszych wykresach. Wariancja i odchylenie standardowe są stosunkowo niewielkie.

Jednocześnie współczynnik zmienności wskazuje na trochę mniejszą niż przeciętna zmienność danych, a kurtoza jest mniejsza niż wartość dla rozkładu normalnego. Pozwala to wnioskować, że dane wykazują małą zmienność, a wartości odstających jest bardzo niewiele. Podobnie jak w przypadku danych dotyczących zaróbków współczynnik skończości jest dodatni, jednak w przybliżeniu równy zero, co pozwala sądzić że rozkład danych jest delikatnie prawostronnie skośny, jednak generalnie symetryczny.

Następnie obliczyliśmy wartości średnich: arytmetycznej, geometrycznej i harmonicznej (Tabela 6). Wyniki tych obliczeń umieściliśmy poniżej:

Średnia	Wartość [miesiąc]
Arytmetyczna	26.198820
Geometryczna	24.46708712071423
Harmoniczna	21.68199609549079

Tabela 6: Średnie dla doświadczenia.

Jak możemy zauważyć, średnie również oscylują w okolicy wartości mediany.

3 Analiza zależności liniowej

W poniższej części będziemy analizować zależność liniową zarobków względem doświadczenia. W tym celu wykorzystamy teoretyczny model regresji liniowej dany wzorem:

$$y_i = b_1 x_i + b_0 + \epsilon_i$$

gdzie:

- y_i - zmienna objaśniana,
- x_i - zmienna objaśniająca,
- ϵ_i - błąd,
- b_0, b_1 - parametry modelu.

Dodatkowo model ten zakłada, że:

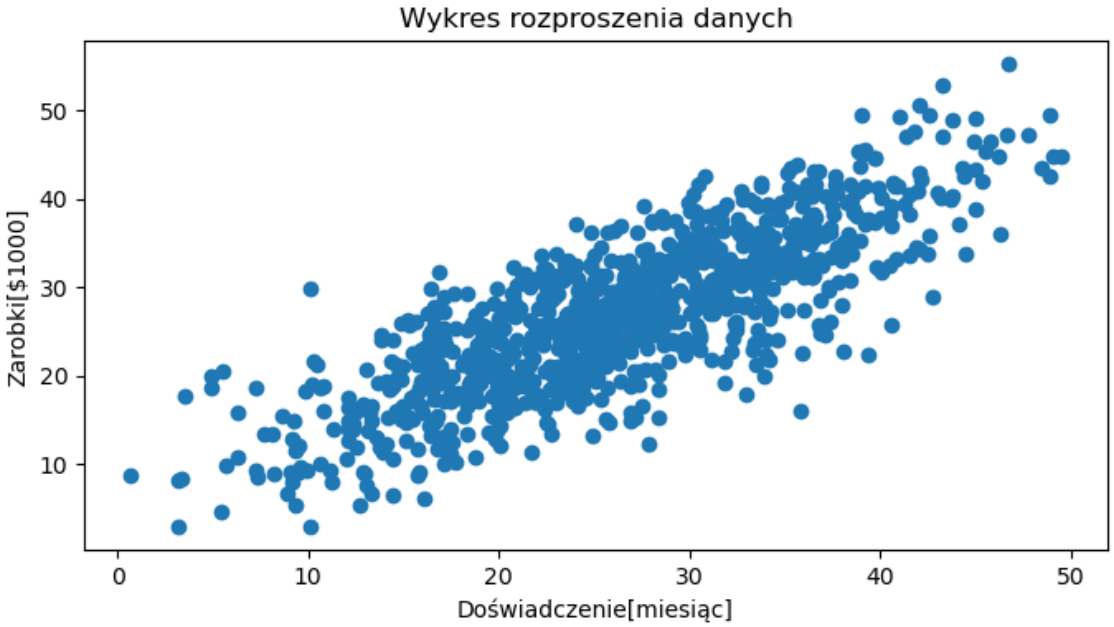
1. $\forall_{i=1,..,n}\ E\epsilon_i = 0$
2. $\forall_{i=1,..,n}\ Var\epsilon_i = \sigma^2$
3. $\epsilon_1, ..., \epsilon_n$ - zmienne losowe, nieskorelowane,
4. $\forall_{i=1,..,n}\ \epsilon_i \sim N(0, \sigma^2)$.

W dalszych obliczeniach przyjęliśmy oznaczenia:

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
- $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$,
- $\hat{y_i} = \hat{b_1} x_i + \hat{b_0}$,
- $n = 988$.

3.1 Wykres rozproszenia

Analizę rozpoczęliśmy od stworzenia wykresu rozproszenia, który umieściliśmy poniżej (Rysunek 15).

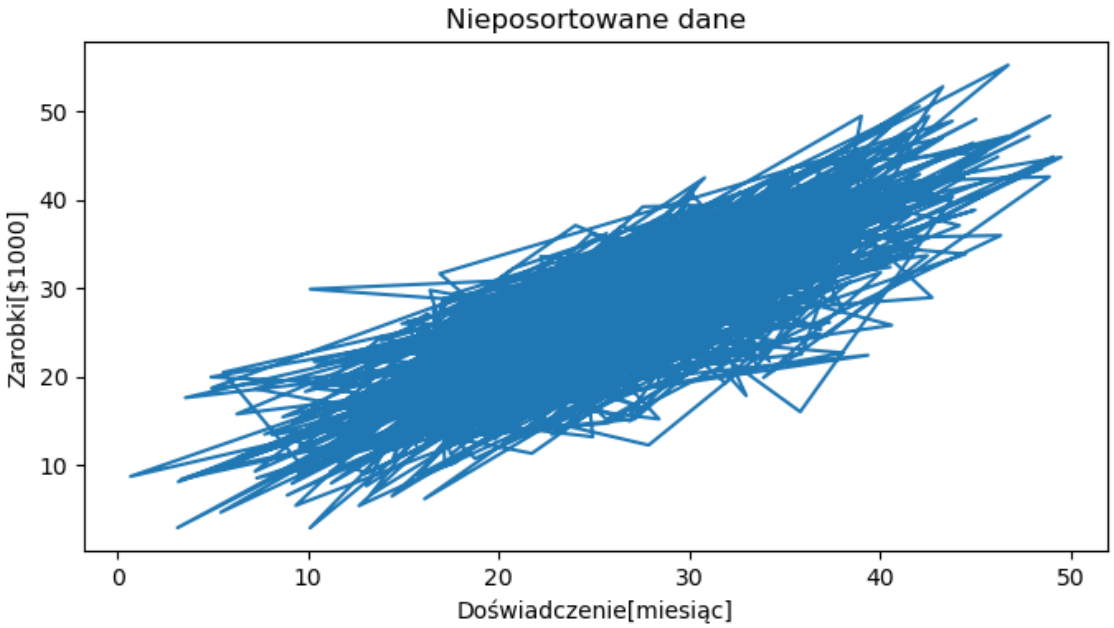


Rysunek 15: Wykres rozproszenia danych

Na powyższym wykresie (Rysunek 15) widzimy pewną tendencję. Dane kumulują się w okolicy punktu (25, 25) który odpowiada medionom dla doświadczenia i zarobków. Oddalając się od tego punktu w kierunku punktów (10, 10) i (50, 50) widzimy, że danych jest coraz mniej. Jednak generalnie możemy zaobserwować że gromadzą się one wokół prostej

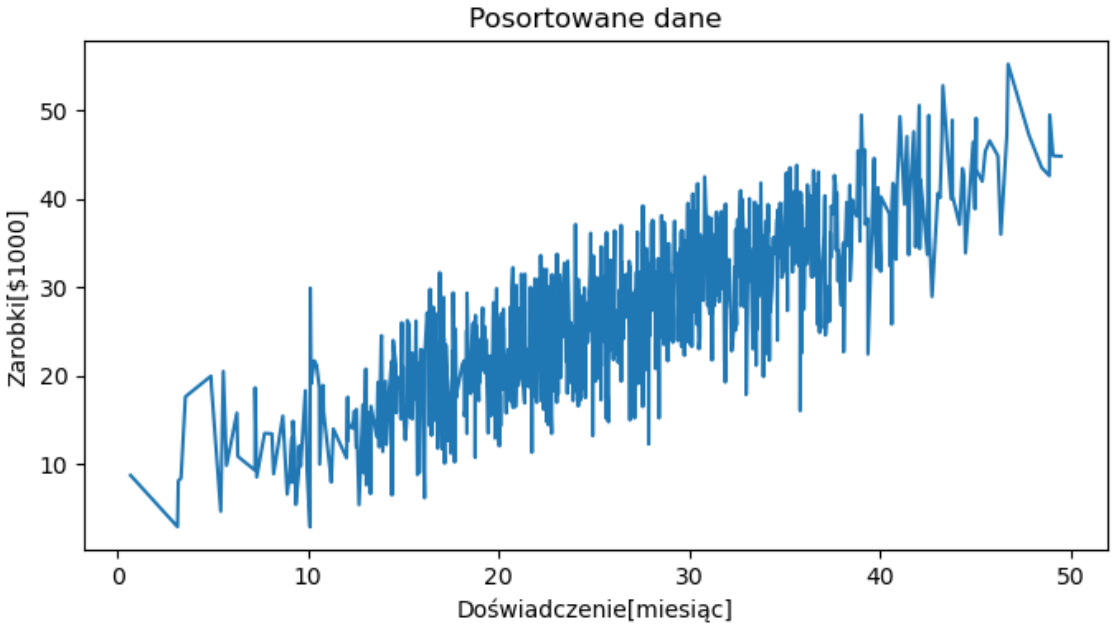
$$y = b_1 x + b_0$$

Zdecydowanie lepiej widać to na wykresie rozproszenia z posortowanymi danymi (Rysunek 17).



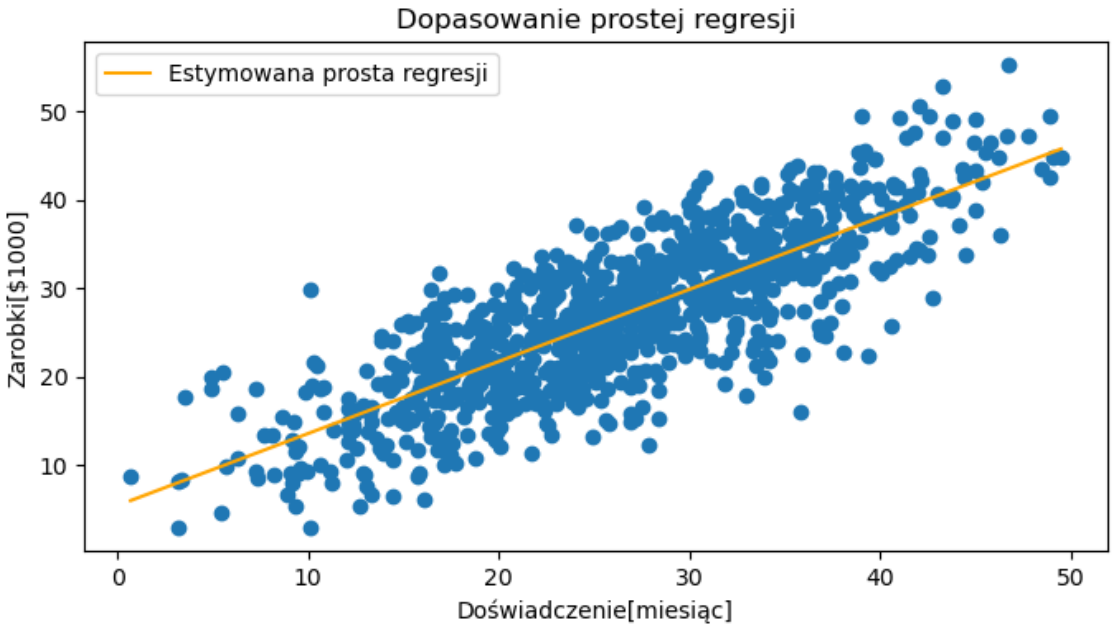
Rysunek 16: Nieposortowane dane

W tym momencie również z posortowanych danych wydzieliliśmy ostatnich 10, które będziemy traktować jako dane testowe i w następnej sekcji dokonamy dla nich predykcji. Teraz jednak zajmiemy się



Rysunek 17: Posortowane dane

dopasowaniem prostej regresji do 988 pierwszych danych. W kolejnych częściach jako zmienną niezależną - x - będziemy traktować dane dotyczące doświadczenia, natomiast jako zmienną zależną - y - dane dotyczące zarobków.



Rysunek 18: Dopasowanie prostej regresji

3.2 Estymacja punktowa

Na początku za pomocy metody najmniejszych kwadratów otrzymaliśmy estymatory $\hat{b_1}$ oraz $\hat{b_0}$ dane następującymi wzorami:

$$\begin{cases} \hat{b_1} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}, \\ \hat{b_0} = \bar{y} - \hat{b_1}\bar{x}. \end{cases}$$

W ten sposób otrzymaliśmy wartości:

$$\begin{cases} \hat{b_1} = 0.8172953853727468, \\ \hat{b_0} = 5.365328792881066. \end{cases}$$

3.3 Estymacja przedziałowa

Dalej dla otrzymanych wyżej estymatorów chcielibyśmy wyznaczyć przedziały ufności na poziomie ufności 1-α. Do tego celu będziemy potrzebowali wartości oczekiwanych i wariancji rozważanych wyżej estymatorów. Wynoszą one odpowiednio dla $\hat{b_1}$:

$$\begin{cases} E\hat{b_1} = b_1, \\ Var\ \hat{b_1} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}, \end{cases}$$

oraz dla $\hat{b_0}$:

$$\begin{cases} E\hat{b_0} = b_0, \\ Var\ \hat{b_1} = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}). \end{cases}$$

Jednak my nie znamy teoretycznej wartości σ^2, więc będziemy używać estymatora wariancji danego wzorem:

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{n-2}.$$

Dzięki powyższym statystykom dokonaliśmy standaryzacji estymatorów $\hat{b_1}$ oraz $\hat{b_0}$ otrzymując:

$$\begin{cases} T = \frac{\hat{b_1}-b_1}{\sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \sim T_{n-2}, \\ T = \frac{\hat{b_0}-b_0}{\sqrt{s^2(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \sim T_{n-2}. \end{cases}$$

Teraz wyznaczyliśmy przedział ufności dla parametru b_1 na poziomie ufności 1-α:

$$P(-t_{n-2,1-\frac{\alpha}{2}} \leqslant T \leqslant t_{n-2,1-\frac{\alpha}{2}}) = 1-\alpha.$$

Po podstawieniu statystyki T oraz przekształceniach otrzymaliśmy:

$$P\left(\hat{b_1} - t_{n-2,1-\frac{\alpha}{2}}\frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \leqslant b_1 \leqslant \hat{b_1} + t_{n-2,1-\frac{\alpha}{2}}\frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right) = 1-\alpha.$$

Analogicznie dla parametru b_0:

$$P\left(\hat{b_0} - t_{n-2,1-\frac{\alpha}{2}}\sqrt{s^2(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \leqslant b_0 \leqslant \hat{b_0} + t_{n-2,1-\frac{\alpha}{2}}\sqrt{s^2(\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right) = 1-\alpha.$$

W ten sposób otrzymaliśmy przedziały ufności na poziomie ufności $1-\alpha$ dla parametrów b_0, b_1 dla $\alpha = 0.05$ wynoszące odpowiednio:

$$P(-0.20251239815905409 \leqslant b_1 \leqslant 1.8371031689045476) = 0.95,$$

$$P(4.345521009349264 \leqslant b_0 \leqslant 6.385136576412867) = 0.95.$$

3.4 Jakość dopasowania

Dalej sprawdzaliśmy jakość dokonanego wyżej dopasowania do danych prostej regresji.

W tym celu obliczyliśmy statystyki dane wzorami:

- $SSE = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$ - suma błędów kwadratów,
- $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ - całkowita suma kwadratów,
- $SSR = \sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2$ - regresyjna suma kwadratów,
- $r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$ - współczynnik korelacji Pearson'a,
- $R^2 = \frac{SSR}{SST}$ - współczynnik determinacji.

Otrzymaliśmy następujące wyniki:

- $SSE = 26510.449172264423$,
- $SST = 76646.52482577649$,
- $SSR = 50136.075653512074$,
- $r = 0.8087772642901232$,
- $R^2 = 0.6541206632326159$.

Analizując wartości otrzymanych statystyk możemy dojść do kilku wniosków. Po pierwsze współczynnik korelacji Pearson'a , który jest stosunkowo blisko 1 sugeruje, że dane są zależne liniowo, jednak nie jest to całkowicie liniowa zależność. Współczynnik determinacji swoją wartością jest zdecydowanie dalej od jedynki, co równie sugeruje, żę zależność liniowa nie jest idealna. Potwierdzają to również wartości błędów, które są kilka rzędów większe od danych.

4 Predykcja

W tej części będziemy dokonywać predykcji dla wcześniej przyjętych przez nas danych testowych. W obliczeniach wykorzystywaliśmy obliczone dla analizowanych wcześniej danych testowych wartości estymatorów: $\bar{x}, \bar{y}, \hat{b_0}$ oraz $\hat{b_1}$.

4.1 Predykcja punktowa

W celu dokonania predykcji punktowej będziemy chcieli obliczyć wartość oczekiwaną $\hat{y_0}$ dla x_0:

$$E\hat{y_0} = E(\hat{b_1}x_0 + \hat{b_0}) = E(\hat{b_1})x_0 + E\hat{b_0} = b_1x_0 + b_0.$$

4.2 Predykcja przedziałowa

Teraz z kolei będziemy chcieli wyznaczyć przedziały ufności dla y_0 na poziomie ufności $1 - \alpha$. W tym celu zdefiniowaliśmy statystykę:

$$A = \hat{y_0} - y_0.$$

Następnie obliczyliśmy dla niej wartość oczekiwaną i wariancję:

$$EA = E(y(\hat{x_0}) - y(x_0)) = E(\hat{b_1}x_0 + \hat{b_0}) - E(b_1x_0 + b_0) = b_1x_0 + b_0 - (b_1x_0 + b_0) = 0,$$

$$VarA = Var(\hat{y}(x_0) - y(x_0)) = Var(y(\hat{x_0})) + Var(y(x_0)) = \sigma^2 + \sigma^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}),$$

$$VarA = \sigma^2(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}).$$

W powyższych obliczeniach wykorzystaliśmy poniższe fakty:

- $\hat{y_0}$ i y_0 są niezależne ,
- $Var\ \hat{y_0} = \sigma^2(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2})$,
- $Var\ y_0 = \sigma^2$.

Jednak ze względu na fakt, że nie znamy wartości teoretycznej σ^2, będziemy korzystać ze zdefiniowanego wcześniej estymatora wariancji w celu wyznaczenia przedziałów ufności. Na początku zdefiniowaliśmy statystykę:

$$T = \frac{\hat{y}(x_0) - y(x_0)}{\sqrt{s^2(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2})}} \sim T_{n-2}.$$

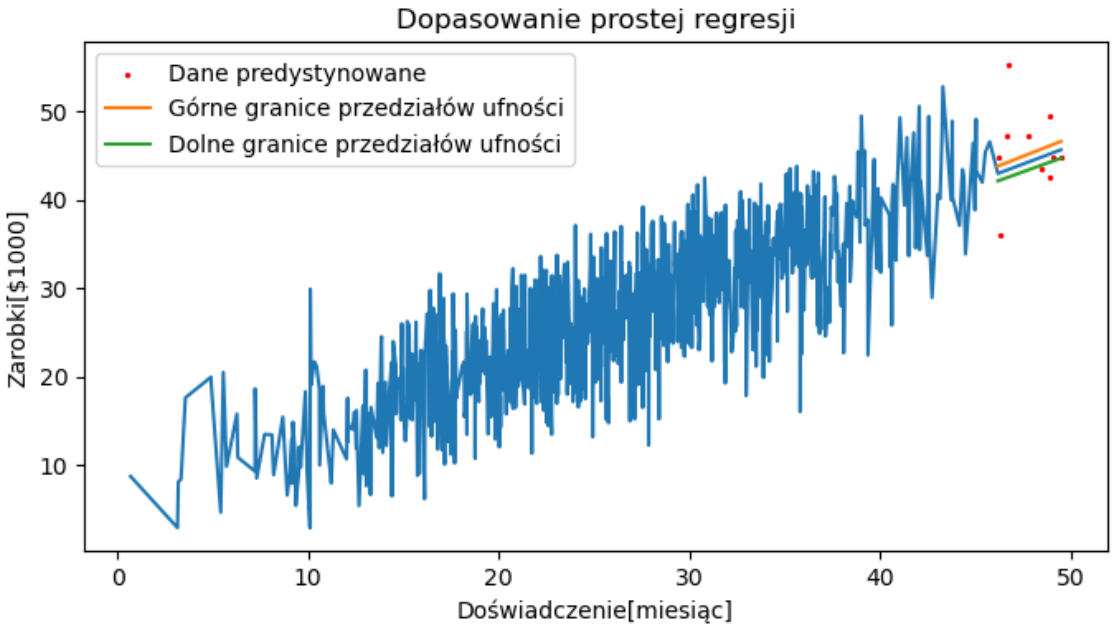
Następnie za jej pomocą wyznaczyliśmy przedział ufności dla estymatora A na poziomie ufności $1-\alpha$:

$$P(-t_{n-2,1-\frac{\alpha}{2}} \leqslant T \leqslant t_{n-2,1-\frac{\alpha}{2}}) = 1-\alpha.$$

Podstawiając i przekształcając otrzyamliśmy:

$$P\left(\hat{y_0} - t_{n-2,1-\frac{\alpha}{2}}\sqrt{s^2(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2})} \leqslant y_0 \leqslant \hat{y_0} + t_{n-2,1-\frac{\alpha}{2}}\sqrt{s^2(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2})}\right) = 1-\alpha.$$

W ten sposób otrzymaliśmy punktową i przedziałową prognozę dla danych, której wyniki zamieściliśmy na poniższym wykresie (Rysunek 19).



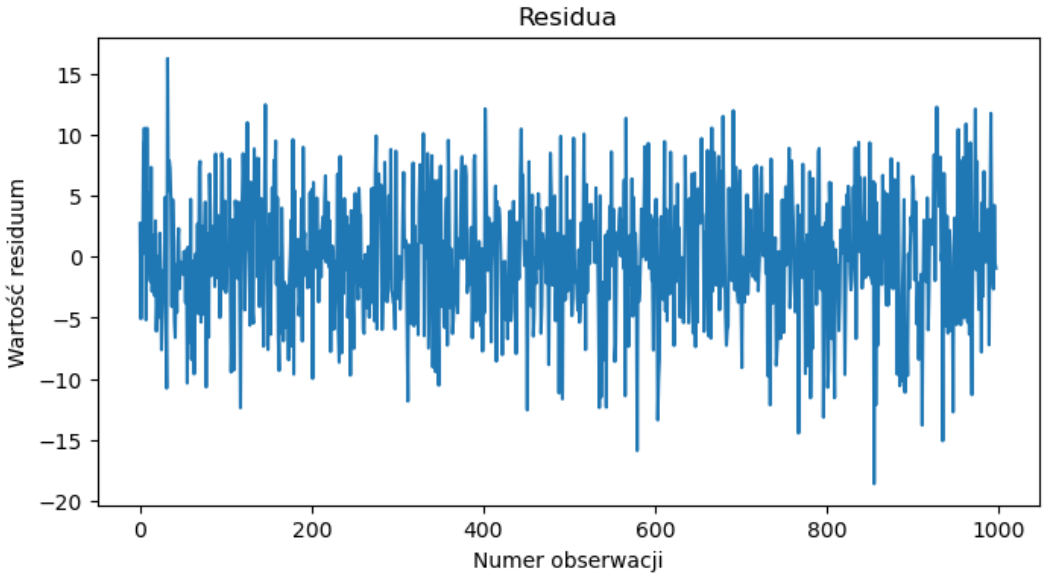
Rysunek 19: Predykcja danych

5 Analiza residuów

W ostatniej części dokonaliśmy analizy residuów, zdefiniowanych jako:

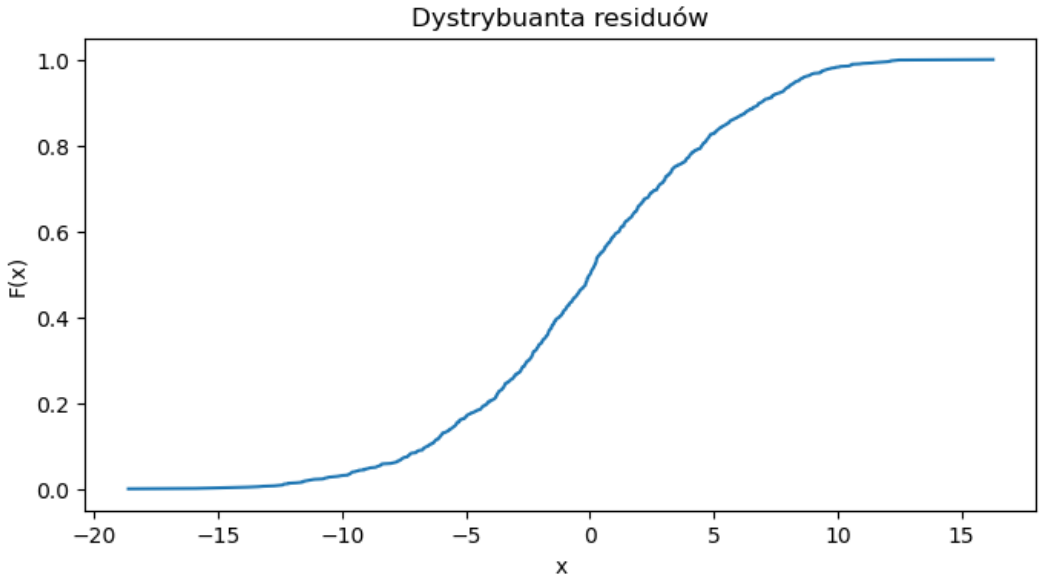
$$e_i = \hat{y_i} - y_i$$

Sprawdzaliśmy, czy spełniają one założenia modelu teorytycznego zdefiniowane w sekcji 3. Na początku narysowaliśmy wykres wartości residuum w zależności od numeru obserwacji (Rysunek 20), w celu sprawdzenia zachowania błędów. Widzimy, że wstępna analiza pozwala przypuszczać, że założenia 1) oraz 2) są spełnione, ponieważ wartości residuów oscylują w okolicy wartości 0 i na całej długości próby odchylają się o mniej więcej podobną wartość. Założenie 1) potwierdza również obliczona wartość średniej równa $1.187560404055959 \cdot 10^{-14}$. Z kolei obliczona przez nas wartość wariancji wynosi 26.563576324914273. Dodatkowo w celu sprawdzenia założenia 2) dokonaliśmy testu Levene'a na jednorodność wariancji. Jego statystyka wyniosła 239.53054966298478, natomiast P-value $4.131453896520822 \cdot 10^{-51}$. Pozwala to stwierdzić, że wariancja w analizowanych przez nas residuach nie jest homogeniczna.



Rysunek 20: Residua

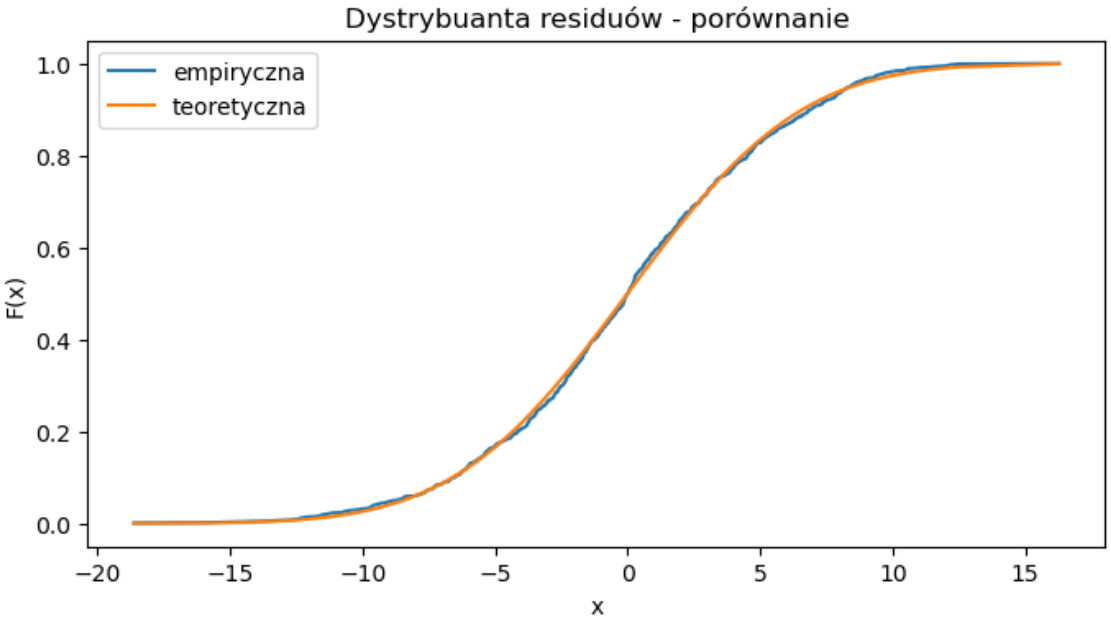
Następnie narysowaliśy dystrybuantę, histogram i gęstość rozkładu residuów (Rysunek 21 i 23) i porównaliśmy je z funkcjami teorytycznymi rozkładu normalnego o wyliczonych wcześniej parametrach (Rysunek 22 i 24). Jak możemy zauważyć, ich wykresy są do siebie zbliżone. Tym samym ostatnim krokiem sprawdzania normalności residuów a tym samym założenia 4) były testy statystyczne. Otrzymaliśmy następujące wyniki (Tabela 7):



Rysunek 21: Dystrybuanta residuów

Test	Statystyka	P-value
Shapiro-Wilka	0.9971351623535156	0.07209804654121399
Jarque-Bera	3.207254193183424	0.20116554627625077

Tabela 7: Testy statystyczne

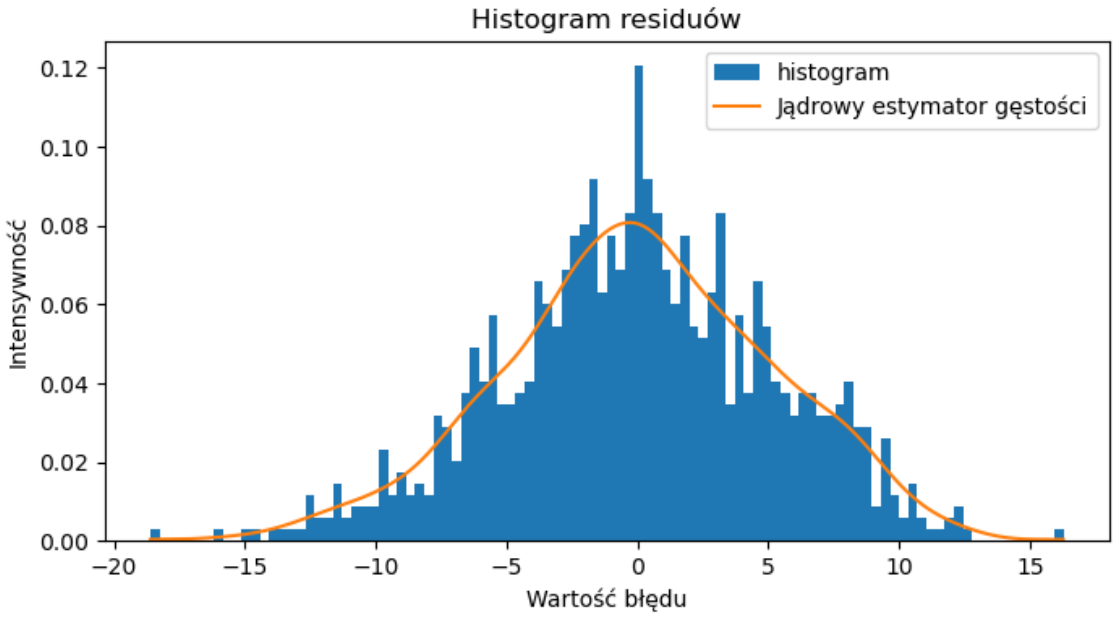


Rysunek 22: Dystrybuanta residuów

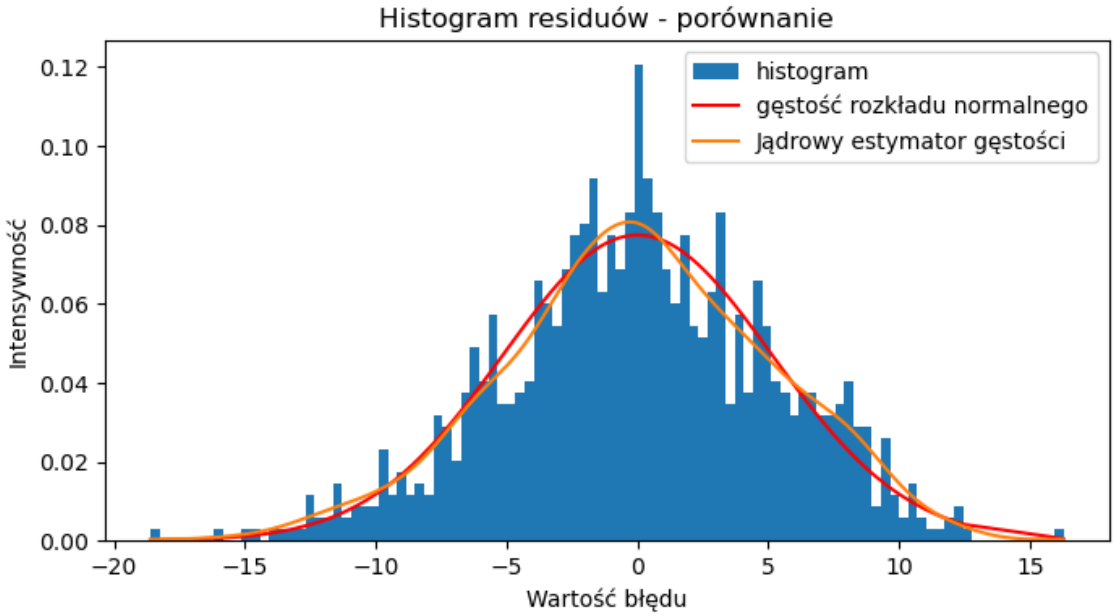
Jako ostatnie sprawdzaliśmy założenie 3) dotyczące nieskorelowania residuów. W tym celu sporządziliśmy wykres autokorelacji próbkowej danej wzorem:

$$Corr(X_t, X_{t+h}) = \frac{\frac{1}{n-h}\sum_{i=1}^{n}(X_t - \bar{X})(X_{t+h} - \bar{X})}{\frac{1}{n}\sum_{i=1}^{n}(X_t - \bar{X})^2}$$

Wyniki umieściliśmy na poniższym wykresie (Rysunek 25).

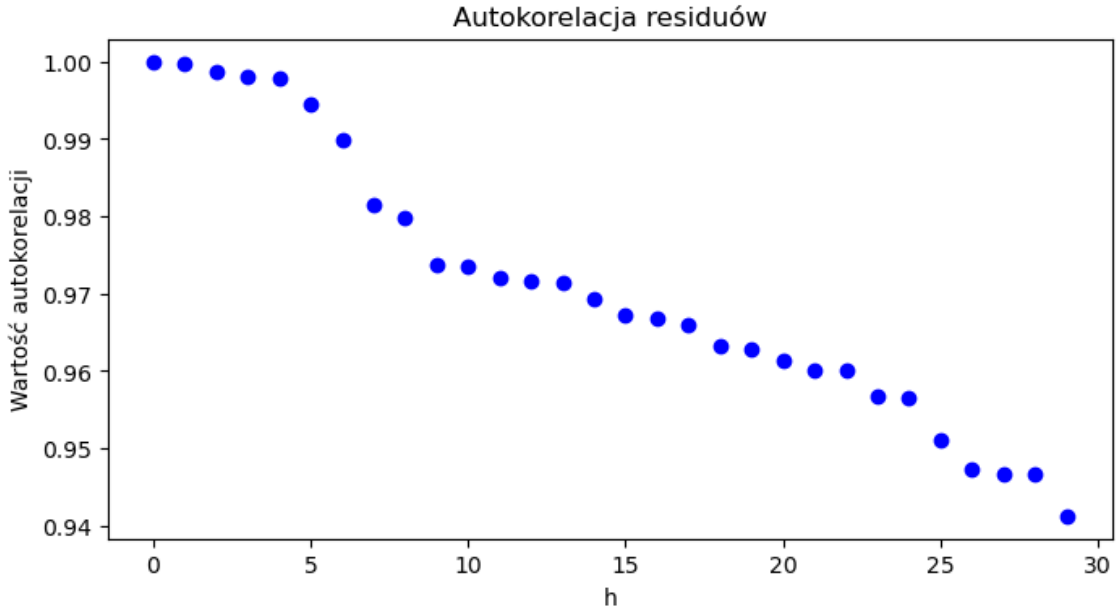


Rysunek 23: Histogram residuów



Rysunek 24: Histogram residuów - porównanie

Jak widzimy wykres autokorelacji w zależności od h wygląda inaczej, niż moglibyśmy się tego spodziewać. Możemy zauważyć, że w miarę wzrostu h, wartość autokorelacji maleje. Jednak jest to bardzo niewielka zmiana, a wartosc funkcji dla $h = 30$ różni się od 1 tylko o 0.05. Może to świadczyć o tym, że residua są skorelowane, a w danych, oprócz zależności liniowej, ukrywa się dodatkowa tendencja. również podobny wniosek możemy wyciągnąć z wyników testu na jednorodność wariancji.



Rysunek 25: Residua - autokorelacja

6 Podsumowanie

Podsumowując jednowymiarową analizę, zarówno dla danych dotyczących zarobków jak i doświadczenia możemy dojść do podobnych wniosków. Cała przeprowadzona przez nas analiza doprowadziła nas do konkluzji, że dane pochodzą z rozkładu normalnego. Przeprowadzone przez nas testy statystyczne również nie odrzucają takiej hipotezy.

Odnośnie analizy zależności liniowej zdecydowanie możemy stwierdzić, że w danych występuje zależność liniowa. Potwierdza to współczynnik korelacji Pearson'a. Jednak już współczynnik determinacji sugeruje, że nie jest to idealnie dopasowany model. Potwierdzają to wartości błędów SSE, SST i SSR, które miały bardzo wysokie wartości w odniesieniu do danych. Takie same wnioski nasuwają się patrząc na wykres rozproszenia. Ewidentnie widac tendencję liniową, jednak dane odchylają się od prostej o znaczące wartości.

Również przyglądając się dokonanej przez nas predykcji widzimy, że prognozowana przez nas prosta dobrze oddaje ogólną tendencję danych, ale tylko jedna na 10 wartości testowych wpadła w wyznaczony przez nas przedział ufności. Wprawia to w wątpliwość fakt, że jest to model odpowiedni dla rozważanych danych.

Ostateczne potwierdzenie naszych przypuszczeń znajdujemy przy analizie residuów. Na pierwszy rzut oka możemy przypuszczać, że spełniają one założenia teoretyczne modelu. Potwierdzają to testy na normalność ich rozkładu i analiza założenia dotyczącego średniej. Jednak przy sprawdzaniu założeń dotyczących wariancji i nieskorelowania residuów doszliśmy do wniosków, że są one skorelowane, a wariancja zmienia się na długości próby. Najprawdopobniej wynika to z faktu, że w danych ukryta jest dodatkowa zależność, której nie uwzględnialiśmy w tym modelu.

Zatem podsumowując model regresji liniowej otrzymany metodą najmniejszych kwadratów dobrze odwzorowuje pewną zależność występującą w danych, jednak nie jest to model w pełni odwzoruwujący dane. Prawdopodobnie, żeby ulepszyć to odwzorowanie powinniśmy dodać w modelu kolejną tendencję, być może okresową.