Komputerowa Analiza Szeregów Czasowych – Raport I

Temat raportu: Analiza Danych z Wykorzystaniem Regresji Liniowej

Nazwisko i Imię prowadzącego kurs:

Wykonawca:	
Imię i Nazwisko, nr indeksu, wydział:	Bartłomiej Brzozowski 268746 Arkadiusz Horodecki 268819 Wydział Matematyki
Termin zajęć:	Poniedziałek, 17:05
Numer grupy ćwiczeniowej:	Grupa 2
Data oddania sprawozdania:	20.12.2023
Ocena końcowa	

1.Wprowadzenie.

1.1 Cele Raportu.

- Utrwalenie wiedzy z zakresu opisu danych, wykonywaniu wykresów i analizy podstawowych statystyk opisowych.
- Nabranie umiejętności w poszukiwaniu danych.
- Dobranie prostej regresji do danych i wyznaczenie współczynnika R^2.
- Wyznaczenie przedziałów ufności dla parametrów modelu regresji liniowej: β_0 i β_1.
- Przeprowadzenie analizy residuów.
- Stworzenie prognozy przyszłej wartości $Y(x_0)$.
- Wyciągnięcie wniosków z otrzymanych wyników.

1.2 Opis Danych.

Przeprowadzono analizę danych dotyczących cen akcji po otwarciu giełdy dla przedsiębiorstwa Amazon (zmienna objaśniająca) i dystrybutora filmów Netflix (zmienna objaśniana) od dnia 23.05.2002 do 05.12.2023. Badano próbki o długościach 5423 (giełda jest zamknięta w święta, soboty i niedziele), z dwoma zmiennymi w każdym przypadku:

- Cena akcji po otwarciu giełdy (Open),
- Data (Date) – dzień w którym akcja miała podaną cenę.

Dane są częścią większego ich zbioru.

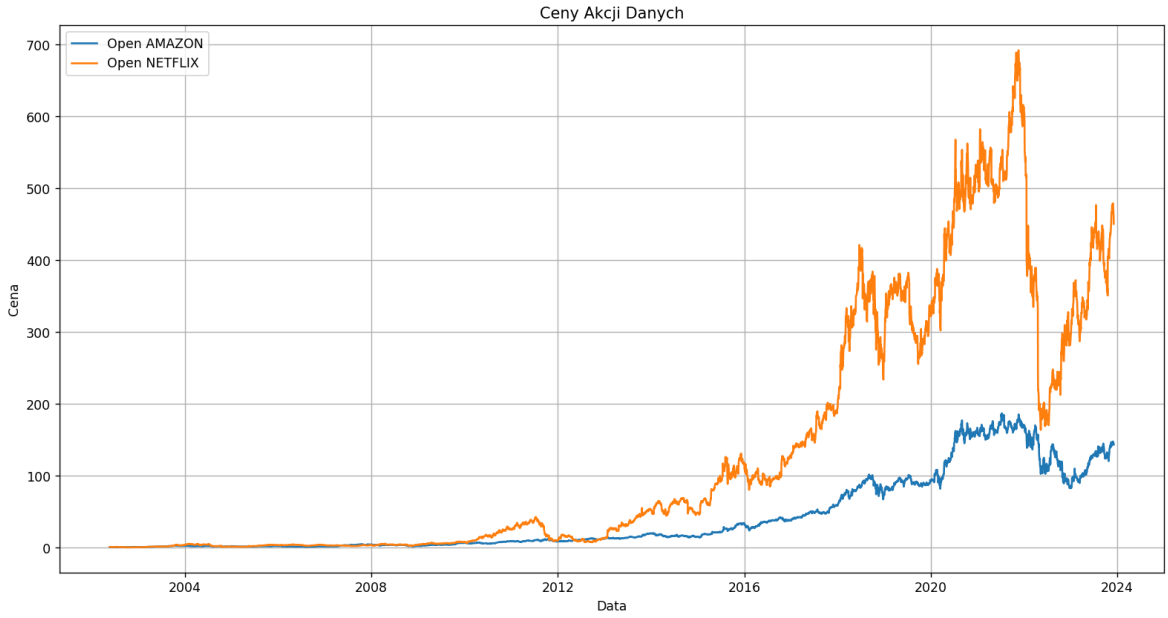
(Dane pochodzą ze strony: https://www.kaggle.com)

1.3 Dane.

W poniższej tabeli (Rys.1), w każdym z wierszy, zaprezentowano przykładowe dane do przeprowadzonej analizy. Kolejno na wykresie przedstawiono dane (Rys.2).

Data	Open Amazon	Open Netflix
23.05.2002	0.942	1.156
24.05.2002	0.970	1.214
28.05.2002	0.979	1.214
29.05.2002	0.941	1.164
30.05.2002	0.935	1.108

Rysunek 1 – Przykładowe dane



Rysunek 2 - Wykres danych.

2. Analiza Jednowymiarowa Danych.

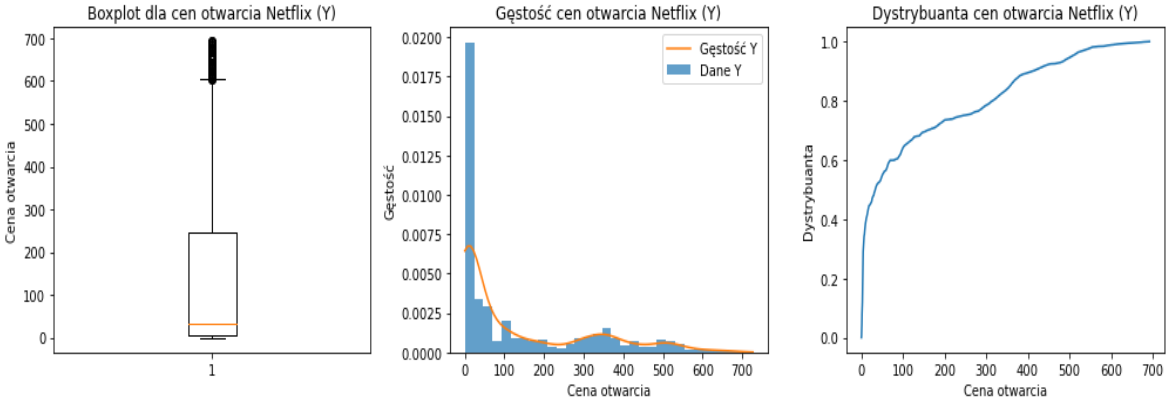
2.1 Zmienna Objaśniająca (X).

Poniżej przedstawiono dane objaśniające za pomocą trzech wykresów (Rys.3). Box-plot przedstawia rozproszenie danych zmiennej, drugi z wykresów przedstawia histogram gęstości zmiennej, a ostatni jej empiryczną dystrybuantę. Dla przedstawionych danych dalej obliczono podstawowe miary położenia, rozproszenia, skośności oraz spłaszczenia, których przybliżone wartości zapisano w tabeli w sekcji 2.4 (Rys.6).

Rysunek 3 - Wykresy: box-plot, gęstości i dystrybuanty zmiennej X.

2.2 Zmienna Objaśniana (Y).

Przedstawiono, tak jak w przypadku zmiennej objaśniającej, dane objaśniane za pomocą trzech wykresów. Box-plot przedstawia rozproszenie danych zmiennej, drugi z wykresów przedstawia histogram gęstości zmiennej, a ostatni jej empiryczną dystrybuantę. Dla przedstawionych danych dalej również obliczono podstawowe miary położenia, rozproszenia, skośności oraz spłaszczenia, których przybliżone wartości zapisano w tabeli w sekcji 2.4 (Rys.6).



Rysunek 4 - Wykresy: box-plot, gęstości i dystrybuanty zmiennej Y.

2.3 Wzory i Definicje Statystyk Opisowych.

Do wyliczeń podstawowych statystyk, skorzystano ze wzorów znajdujących się na poniżej grafiki (Rys.5).

- Średnia arytmetyczna: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$;
- Średnia harmoniczna: $\frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$;
- Średnia ucinana: $\frac{1}{n-2k} \sum_{i=k+1}^{n=k} x_i$;
- Średnia Winsorowska: $\frac{1}{n}\left[(k+1)x_{k+1} + \sum_{i=k+2}^{n=k-1} x_{i+(k+1)} \cdot x_{(n-k)}\right]$;
- Wariancja: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$;
- Odchylenie standardowe: $S = \sqrt{S^2}$;
- Współczynnik zmienności: $\vartheta = \frac{1}{\overline{x}} \cdot 100\%$;
- Współczynnik skośności: $\alpha = \frac{n}{(n-1)(n+2)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S}\right)^3$;
- Kurtoza: $K = \frac{n-1}{(n-2)(n-3)}\big((n+1)k - 3(n-1)\big) + 3$;
- Rozstęp: $\text{rozstęp} = \max(x_i) - \min(x_i)$;
- Kwantyle: $x_p = \begin{cases} x_{(np)} & ; np - całkowitego \\ (1 - \lfloor np \rfloor)x_{\lfloor np \rfloor + 1} + \lfloor np \rfloor x_{\lfloor np \rfloor + 2} & ; np - niecałkowitego \end{cases}$;
- Mediana i Kwartyle: $x_{med} = x_{0,5} = Q2$,$x_{0,75} = Q3$ i $x_{0,25} = Q1$;
- Rozstęp międzykwartylowy: $IQR = Q3 - Q1$.

Rysunek 5 - Wzory do statystyk opisowych.

2.4 Wyniki Wielkości Statystycznych dla Danych.

Zaimplementowano w języku Python, wyżej wymienione wzory (Rys.5) do obliczeń podstawowych statystyk opisowych dla cen akcji przedsiębiorstwa Amazon i dystrybutora filmów Netflix. Zapisano wyniki w niżej zamieszczonej tabeli (Rys.6).

Statystyki	Open Amazon	Open Netflix
Średnia arytmetyczna	$\overline{x} = 41.598$	$\overline{y} = 131.304$
Średnia harmoniczna	4.954	6.199
Średnia ucinana	31.921	100.046
Średnia Winsorowska	39.112	121.311
Wariancja	$S_x^2 = 2759.966$	$S_y^2 = 29912.277$
Odchylenie standardowe	$S_x = 52.535$	$S_y = 172.952$
Współczynnik zmienności	$\vartheta_X = 126.292$	$\vartheta_y = 131.718$
Współczynnik skośności	$\alpha_x = 1.234$	$\alpha_y = 1.236$
Kurtoza	$K_x = 3.162$	$K_y = 3.275$
Minimum	$\min(x) = 0.619$	$\min(y) = 0.378$
Maksimum	$\max(x) = 187.200$	$\max(y) = 692.350$
Rozstęp	$rozstęp_x = 186.581$	$rozstęp_y = 691.972$
Q1	$x_{0.25} = 3.183$	$y_{0.25} = 4.108$
Mediana	$x_{med} = 13.139$	$y_{med} = 33.701$
Q3	$x_{0.75} = 81.111$	$y_{0.75} = 244.832$
Rozstęp międzykwartylowy	$IQR_x = 77.927$	$IQR_y = 240.725$

*R*ysunek 6 - Statystyki opisowe dla zmiennych X i Y.

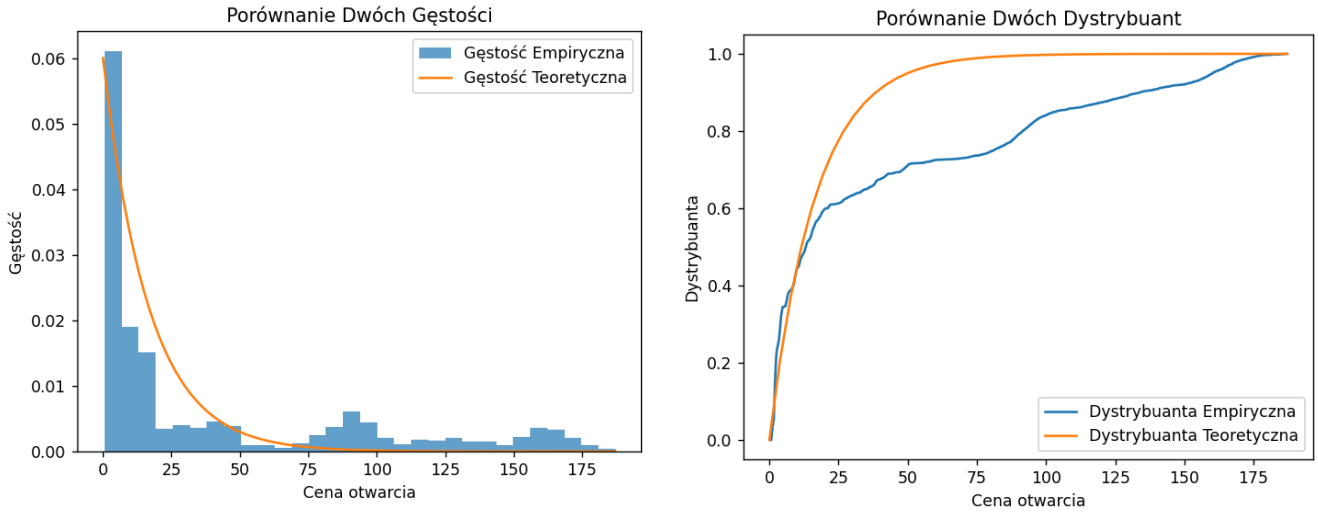
2.5 Interpretacja Wyników i Wnioski.

Z wcześniejszej tabeli (Rys.6), wnioskować możemy, ze średnich arytmetycznych (mówi o środku ciężkości zbioru danych), że większość cen otwarcia akcji firmy Amazon wynosiła około 41.598, a firmy Netflix wynosiła około 131.304. Średnia ucinana w obydwu przypadkach, w szczególności w przedsiębiorstwie Netflix, nie jest zbliżona do wartości średnich arytmetycznych, przez co wnioskować można, że skrajne wartości wywierają duży wpływ na wartości średnich, w przypadku cen otwarcia akcji firmy Netflix potwierdza to

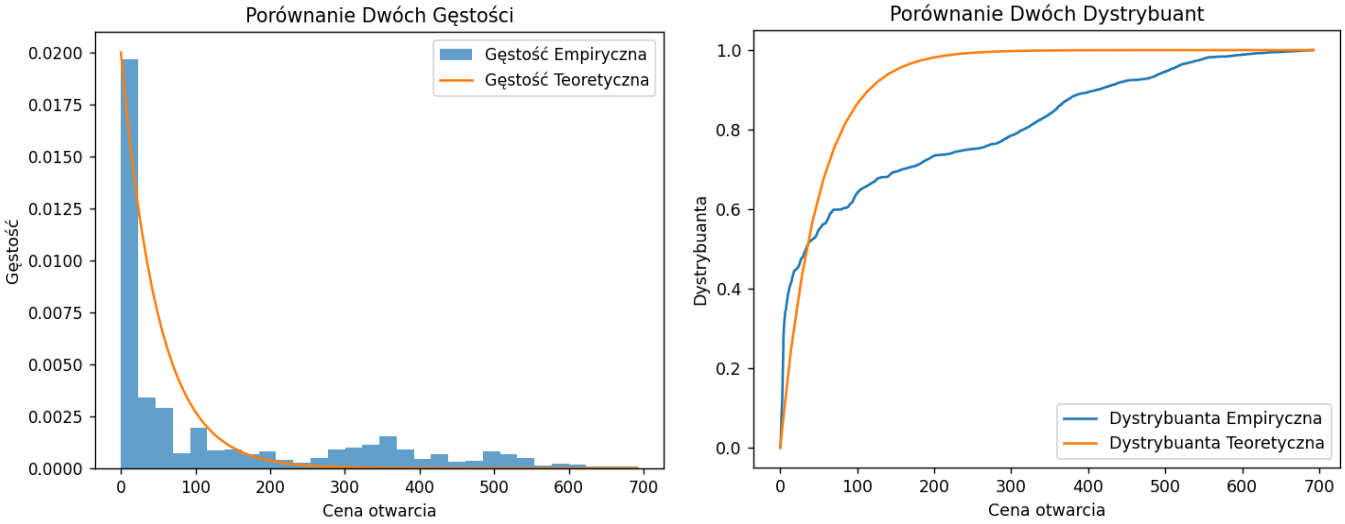
również wykres typu Box-plot (Rys.4). Dalej rozstęp, wartość minimalna i maksymalna, w przypadku firmy Amazon informuje nas o dużym zakresie zmienności, a w przypadku firmy Netflix, o jeszcze większym zakresie. Natomiast współczynniki zmienności, dla obu przedsiębiorstw również sugerują dużą zmienność danych. Kolejno współczynnik skośności, dodatni w obu zbiorach danych, sugeruje, że rozkłady cen są lekko przesunięte w prawo. Oznacza to, że istnieją pewne przypadki z wyższymi cenami otwarcia akcji. Firmy Amazon i Netflix, mają podobne wartości kurtozy (3.162 i 3.275), co wskazuje, że rozkłady cen otwarcia są umiarkowanie leptokurtyczne (stożkowate). Mediana dla przedsiębiorstwa Amazon (13.139) jest znacznie niższa niż średnia, co może wskazywać na występowanie nielicznych przypadków z wyższymi cenami otwarcia. Kwartyle także pokazują szeroki zakres cen. Dla firmy Netflix mediana (33.701) jest wyraźniej mniejsza od średniej, co razem z Kwartylami, wskazuje, na ogólnie wyższe ceny akcji, ale również znaczny zakres zmienności. Dane dla firmy Amazon wyróżniają się mniejszą wariancją (2759.966) i odchyleniem standardowym (52.535), od wariancji (29912.277) i odchylenia standardowego (172.952) danych firmy Netflix, co wskazuje na mniejszą zmienność cen otwarcia akcji przedsiębiorstwa Amazon niż cen otwarcia akcji przedsiębiorstwa Netflix. Wartości średniej harmonicznej i Winsorowskiej, dla obu zbiorów danych, różnią się od siebie, co sugeruje, że rozkład cen otwarcia akcji nie jest idealnie symetryczny, może być bardziej skomplikowany.

Po analizie wykresów dystrybuant i gęstości: cen otwarcia akcji firmy Amazon (Rys.3) oraz cen otwarcia akcji firmy Netflix (Rys.4), nasuwającym się wnioskiem jest podobieństwo rozkładu cen otwarcia obu przedsiębiorstw. Dodatkowo dopasowano do gęstości i dystrybuant empirycznych, znane, najbardziej zbliżone rozkłady. W obu przypadkach wybrano rozkład wykładniczy (z różnymi współczynnikami λ), ponieważ całkowity błąd dopasowania nie jest duży (jest w granicach tolerancji). Dla rozkładu cen

otwarcia akcji firmy Amazon dopasowano rozkład Wykładniczy z parametrem $\lambda = 0.06$ i przedstawiono porównanie wyżej wspomnianych funkcji gęstości i dystrybuanty empirycznej z ich teoretycznymi odpowiednikami (Rys.7). Podobnie dla rozkładu cen otwarcia akcji firmy Netflix dopasowano rozkład wykładniczy z parametrem $\lambda = 0.02$ i również przedstawiono poniżej porównanie wyżej wymienionych funkcji gęstości i dystrybuanty empirycznej z ich teoretycznymi odpowiednikami (Rys.8).



Rysunek 7 - Dopasowanie rozkładu wykładniczego do rozkładu zmiennej X.

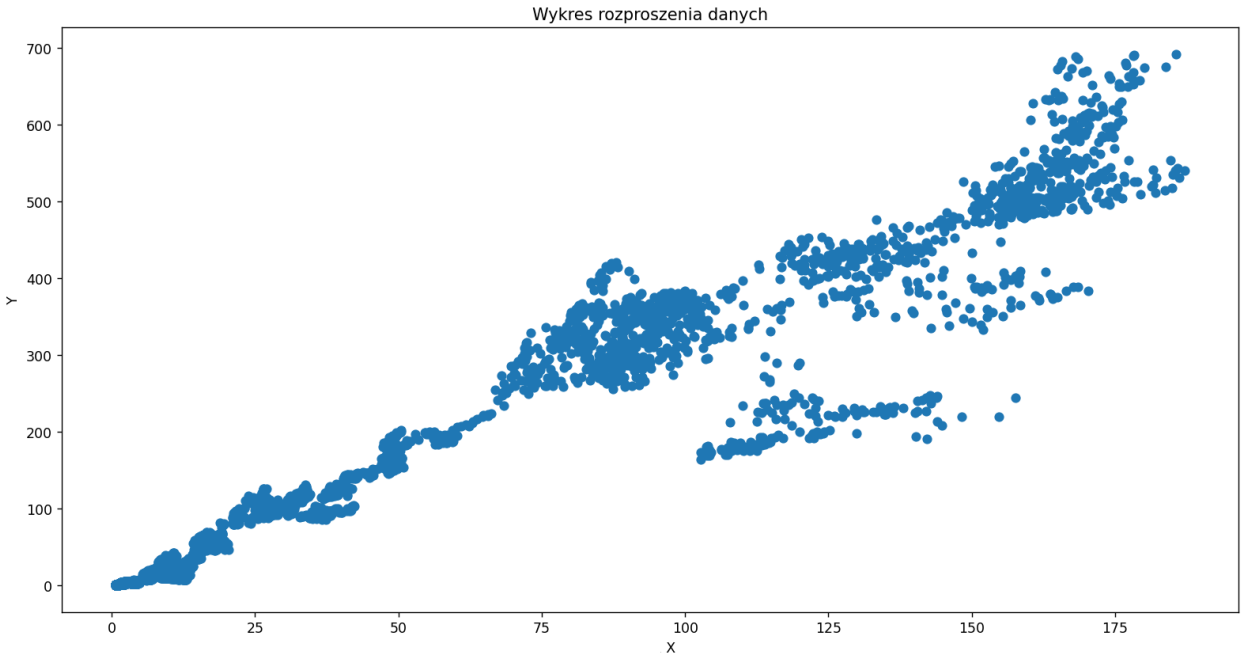


Rysunek 8 - Dopasowanie rozkładu wykładniczego do rozkładu zmiennej Y.

3. Analiza Zależności Liniowej Między Zmienną Zależną i Zmienną Niezależną.

3.1 Wykres Rozproszenia.

Przedstawiono na poniższym wykresie rozproszenia (Rys.9) zmienne, tak jak wspomniano w podpunkcie 1.2. Jako zmienną Objaśniającą X, wzięto cenę otwarcia akcji firmy Amazon, a jako zmienną Objaśnianą Y, wzięto cenę otwarcia akcji przedsiębiorstwa Netflix.



Rysunek 9– Wykres rozproszenia danych.

Zauważalna jest możliwość wystąpienia zależności liniowej między zmiennymi, widoczne jednak są obserwacje odstające.

3.2 Wzory i Definicje.

Do dalszej analizy regresji liniowej skorzystano z poniższych definicji i wzorów (Rys.10 i Rys.11).

- Warunki Metody Najmniejszych Kwadratów:
  1. $\frac{dS}{db_0} = -\sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_i) = 0 \Leftrightarrow \sum_{i=1}^{n} e_i = 0;$
  2. $\frac{dS}{db_1} = \sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_i)(-x_i) = 0 \Leftrightarrow \sum_{i=1}^{n} e_i x_i = 0;$

  Gdzie $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$.

Rysunek 10 – Wzory i definicje do analizy regresji liniowej cz.1.

- Całkowita suma kwadratów: $SST = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$;
- Błędna suma kwadratów: $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$;
- Regresyjna suma kwadratów: $SST = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$;
- Współczynnik R^2: $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$;
- Próbkowy Empiryczny Współczynnik Korelacji:
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2 \sum_{i=1}^{n}(y_i - \bar{y}_i)^2}};$$
- Błąd średniokwadratowy: $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$;
- Teoretyczny Model Regresji Liniowej:
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1,2,3,\dots$$
Gdzie:
    1. $X_1, X_2, \dots, X_n$-wielkości deterministyczne;
    2. β_0, β_1-parametry modelu;
    3. σ^2-parametr modelu.

  Założenia:
    1. $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$-zmienne losowe, które są nieskorelowane;
    2. $\forall_i\ E(\varepsilon_i) = 0$;
    3. $\forall_i Var(\varepsilon_i) = 0$;
    4. $\{\varepsilon_i\}_{i=1}^{n} \rightarrow$ jest białym szumem $WN(0; \sigma^2)$.

  Estymatory: $\hat{\beta}_0\ i\ \hat{\beta}_1$z Metody Najmniejszych Kwadratów:
    1. $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
    2. $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_i)(Y_i - \bar{Y}_i)}{\sum_{i=1}^{n}(X_i - \bar{X}_i)^2}$.
- Rozkłady: $\hat{\beta}_0\ i\ \hat{\beta}_1$z Metody Najmniejszych Kwadratów:
    1. $E\hat{\beta}_0 = \beta_0$ i $Var\hat{\beta}_0 = \sigma^2\left(\frac{1}{n} + \frac{\overline{(x)}^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}\right) \Rightarrow\ \hat{\beta}_0 \sim N(E\hat{\beta}_0, Var\hat{\beta}_0)$;
    2. $E\hat{\beta}_1 = \beta_1$i $Var\hat{\beta}_0 = \left(\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}\right) \Rightarrow\ \hat{\beta}_1 \sim N(E\hat{\beta}_1, Var\hat{\beta}_1)$.
- Przedziały ufności:
    1. σ-nieznana dla parametru β_0:
$$\beta_0 \in \left[\hat{\beta}_0 - t_{n-2,1-\frac{\alpha}{2}} \cdot S \cdot \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}}; \hat{\beta}_0 + t_{n-2,1-\frac{\alpha}{2}} \cdot S \cdot \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}}\right];$$
    2. σ-nieznana dla parametru β_1:
$$\beta_1 \in \left[\hat{\beta}_1 - t_{n-2,1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}}; \hat{\beta}_1 + t_{n-2,1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}}\right].$$

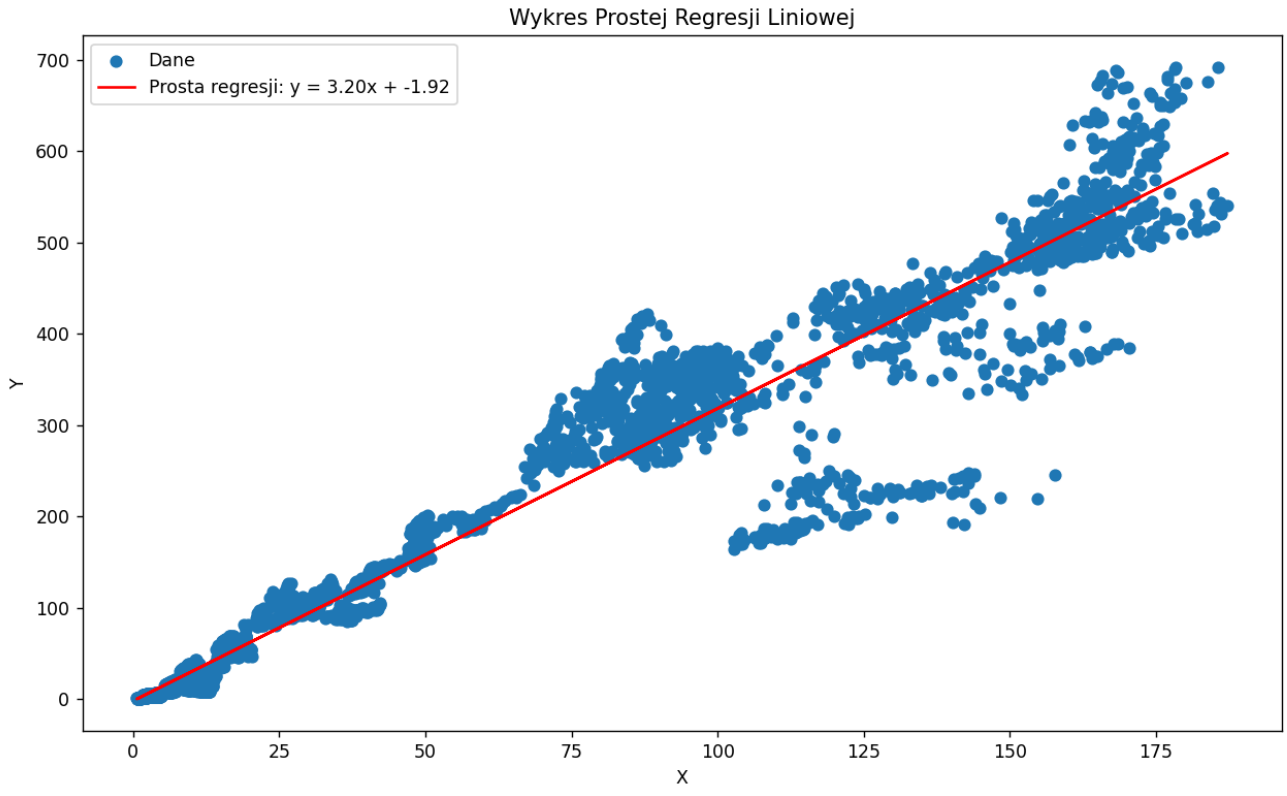
Rysunek 11 - Wzory i definicje do analizy regresji liniowej cz.2.

3.3 Usunięcie Wartości Odstających.

Do usunięcia obserwacji odstających, skorzystano ze wzorów (Rys.10-11) do wyznaczenia prostej regresji w postaci funkcji $y_i = b_1 x_i + b_0$, dla danych początkowych, za pomocą Metody Najmniejszych Kwadratów, otrzymane współczynniki wynosiły:

$$b_1 \approx 3.20$$
$$b_0 \approx -1.92$$

Na poniższym wykresie (Rys.12) zaprezentowano prostą regresje liniową dla danych początkowych (Rys.9), przy wyżej wyznaczonych współczynnikach.

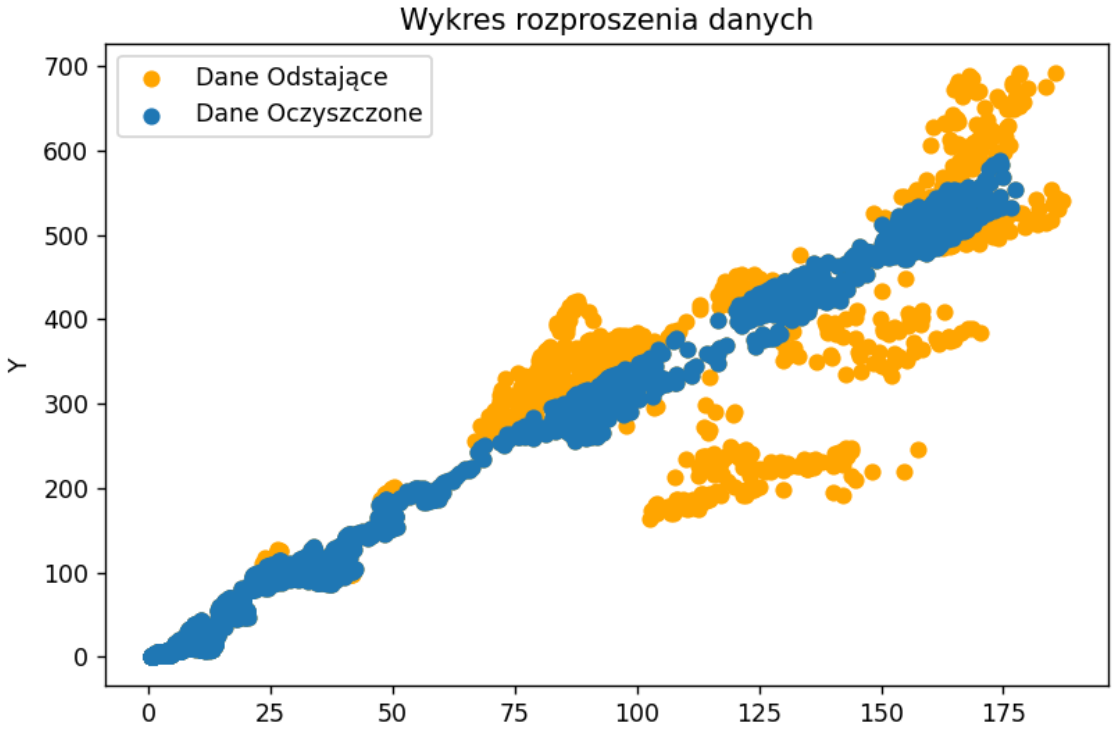


Rysunek 12 – Wykres prostej regresji liniowej dla danych początkowych.

Następnie do usunięcia wartości odstających skorzystano ze wzoru na e_i (Rys.10) i przyjęto warunek, że wszystkie wartości, które go nie spełniają, są odrzucane. Poniżej zapisano wspomniany warunek:

$$Q1 - 1.5 \cdot IQR \leq e_i \leq Q3 + 1.5 \cdot IQR.$$

Kolejno na poniższym wykresie (Rys.13), zaznaczono dane odstające i te, które spełniają, wyżej wspomniany warunek.



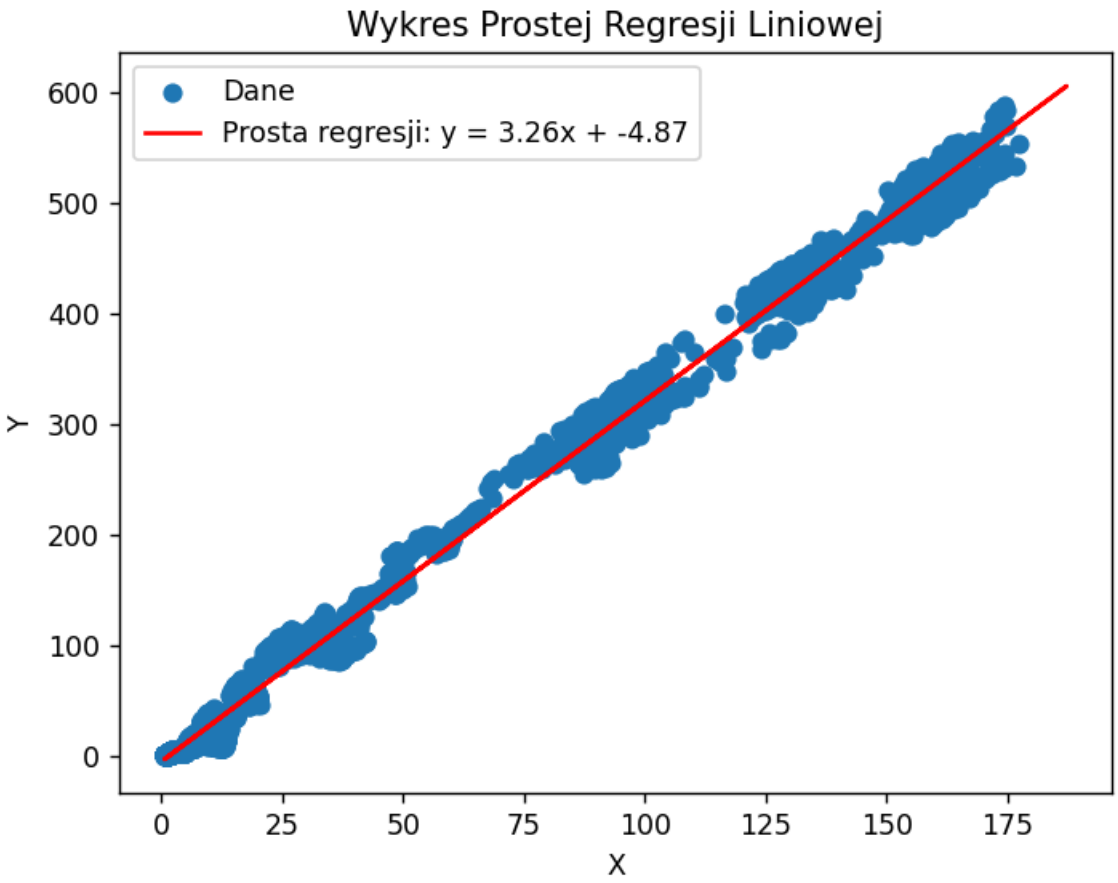
Rysunek 13 –Wykres rozproszenia danych z wyróżnieniem wartości odstających.

3.4 Wyznaczenie współczynników prostej regresji.

Po dokonaniu usunięcia obserwacji odstających (Rys.13), skorzystano ze wzorów (Rys.10-11) do wyznaczenia prostej regresji w postaci funkcji $y_i = \beta_1 x_i + \beta_0$, dla danych oczyszczonych, za pomocą Metody Najmniejszych Kwadratów, otrzymane współczynniki wynosiły:

$$\beta_1 \approx 3.26$$
$$\beta_0 \approx -4.87$$

Na poniższym wykresie (Rys.14) zaprezentowano prostą regresje liniową dla danych oczyszczonych (Rys.13), przy wyżej wyznaczonych współczynnikach.



Rysunek 14 - Wykres prostej regresji liniowej dla danych początkowych.

3.5 Wyznaczenie przedziałów ufności dla β_1.

Do wyznaczenia przedziału ufności dla współczynnika β_1, przy nieznanej σ. Korzystamy, z poniższego faktu:

$$T = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \sim T_{n-2}, \text{ gdzie } S^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{n-2}.$$

A T_{n-2} to rozkład t-Studenta z $n-2$ stopniami swobody, w naszym przypadku n to długość próbki, czyli $n = 4572$ (dane po oczyszczeniu). Co dalej pozawala zapisać warunek na przedział ufności: $P(A \leq \beta_1 \leq B) = 1 - \alpha$, dla małego alfa, jako

$$P\left(-t_{n-2,1-\frac{\alpha}{2}} \leq T \leq t_{n-2,1-\frac{\alpha}{2}}\right) = 1 - \alpha,$$

gdzie $t_{n-2,1-\frac{\alpha}{2}}$ to kwantyl rzędu $1 - \frac{\alpha}{2}$.

Po dokonaniu podstawienia za zmienną T, wyżej wspomnianego faktu i kilku przekształceń, zostawiając β_1 w środku nierówności, otrzymujemy wzór z sekcji 3.2 na przedział ufności dla parametru β_1 (Rys.11).

$$\beta_1 \in \left[\hat{\beta}_1 - t_{n-2,1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x}_i)^2}}; \hat{\beta}_1 + t_{n-2,1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x}_i)^2}}\right].$$

Po przyjęciu, że $\alpha = 0.05$ i podstawieniu wartości numerycznych otrzymujemy:

$$\beta_1 \in [3.11, 3.30].$$

3.6 Wyznaczenie przedziałów ufności dla β_0.

Do wyznaczenia przedziału ufności dla współczynnika β_0, przy nieznanej σ. Korzystamy, z poniższego faktu:

$$T = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x}_i)^2}}} \sim T_{n-2}, \text{ gdzie } S^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}{n-2}.$$

A T_{n-2} to rozkład t-Studenta z $n-2$ stopniami swobody, w naszym przypadku n to długość próbki, czyli $n = 4572$ (dane po oczyszczeniu). Co dalej pozawala zapisać warunek na przedział ufności: $P(A \leq \beta_0 \leq B) = 1 - \alpha$, dla małego alfa, jako

$$P\left(-t_{n-2,1-\frac{\alpha}{2}} \leq T \leq t_{n-2,1-\frac{\alpha}{2}}\right) = 1 - \alpha,$$

gdzie $t_{n-2,1-\frac{\alpha}{2}}$ to kwantyl rzędu $1 - \frac{\alpha}{2}$.

Po dokonaniu podstawienia pod zmienną T, wyżej wspomnianego faktu i przekształceń, zostawiając β_0 w środku nierówności, otrzymujemy wzór z sekcji 3.2 na przedział ufności dla parametru β_0 (Rys.11).

$$\beta_0 \in \left[\hat{\beta}_0 - t_{n-2,1-\frac{\alpha}{2}} \cdot S \cdot \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x}_i)^2}}; \hat{\beta}_0 + t_{n-2,1-\frac{\alpha}{2}} \cdot S \cdot \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x}_i)^2}}\right].$$

Po przyjęciu, że $\alpha = 0.05$ i podstawieniu wartości numerycznych otrzymujemy:

$$\beta_0 \in [-6.91, 3.08].$$

3.6 Badanie zależności między zmiennymi.

Skorzystano ze wzorów znajdujących się w sekcji 3.2 (Rys.11), po czym zapisano wyniki współczynników w poniższej Tabeli (Rys.15).

Całkowita suma kwadratów	$SST = 95278317.76$
Regresyjna suma kwadratów	$SSR = 94640940.09$
Błędna suma kwadratów	$SSE = 637377.68$
Współczynnik R^2	$R^2 = 0.99$
Korelacja Spearmana	$S \approx 0.97$
Korelacja Pearsona	$P \approx 0.99$

Rysunek 15 – Współczynniki zależności między zmiennymi.

3.7 Interpretacja wyników i wnioski.

Z wcześniejszej tabeli (Rys.15), wartości współczynników korelacji Spearmana i Pearsona, mówią o silnej dodatniej korelacji między zmiennymi, co wskazuje, na liniową zależność między nimi, opisaną wcześniejszym równaniem

$$y_i = \beta_0 + \beta_1 x_i.$$

Dodatkowo wysoka wartość współczynnika R^2, wskazuję, na dobrze dopasowaną prostą do danych. Również duża różnica między wartościami SSE i SST, sugeruje, że model dobrze dopasowuje się do danych.

4.Analiza Residuów.

4.1 Wzory i Definicje.

Do przeprowadzenia analizy residuów skorzystano z poniższych definicji i wzorów (Rys.16-17).

- Residuum (reszta, błąd): $e_i = \hat{Y}_i - Y_i$
  Traktowane są jako realizacje zmiennych losowych ε_i w modelu teoretycznym regresji liniowej (Rys.10).

  Założenia:

  1. $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \rightarrow$ niezależne (nieskorelowane);
  2. $\forall_i\, E(\varepsilon_i) = 0$;
  3. $\forall_i\, Var(\varepsilon_i) = \sigma^2$;
  4. $\forall_i\, \varepsilon_i \sim N(0, \sigma^2)$.

Rysunek 16 - Wzory i Definicje do Analizy Residuów cz.1.

- Studentyzacja rozkładów β_0 i β_1:
  1. Dla β_0: $T = \frac{\widehat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}}} \sim T_{n-2}$;
  2. Dla β_1: $T = \frac{\widehat{\beta}_1 - \beta_1}{S/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \sim T_{n-2}$;

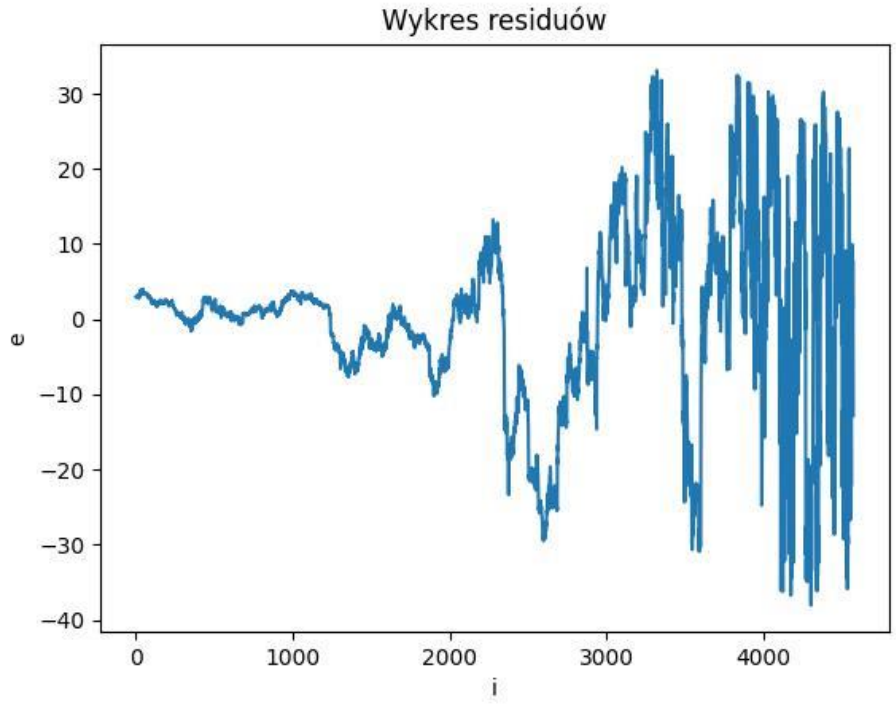
  Gdzie $S^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{n-2}$.
- Normalizacja zmiennej losowej $Z \sim N(\mu, \sigma^2)$:

$$N = \frac{Z - \mu}{\sigma} \sim N(0,1).$$

Rysunek 17 - Wzory i Definicje do Analizy Residuów cz.2.

4.2 Wizualizacja Residuów.

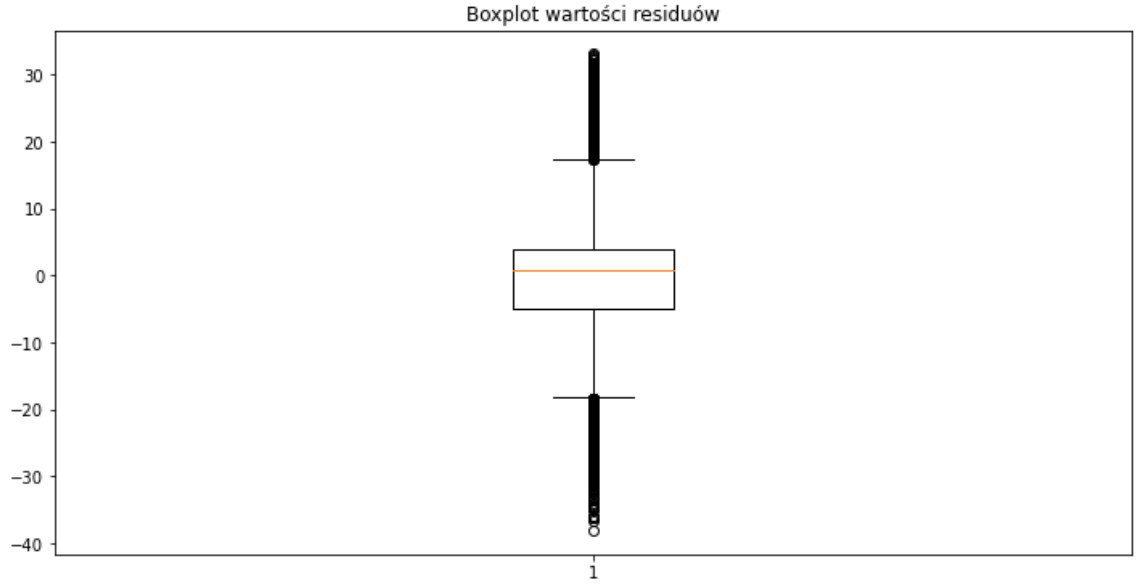
Poniżej przedstawiono wykres residuów od wartości indeksu i (Rys.18).



Rysunek 18 – Wykres rozproszenia residuów.

Interpretując powyższy wykres, możemy, wnioskować, że residua, przez to, że oscylują w okolicy zera, mogą mieć stałą wartość oczekiwaną równą zero: $E(\varepsilon_i) = 0\ \forall_i$, ponadto możemy, wnioskować, że nie mają one stałej wariancji. Warto zauważyć, brak powtarzalności, na wykresie.

Następnie przedstawiono wykres pudełkowy residuów (Rys.19).



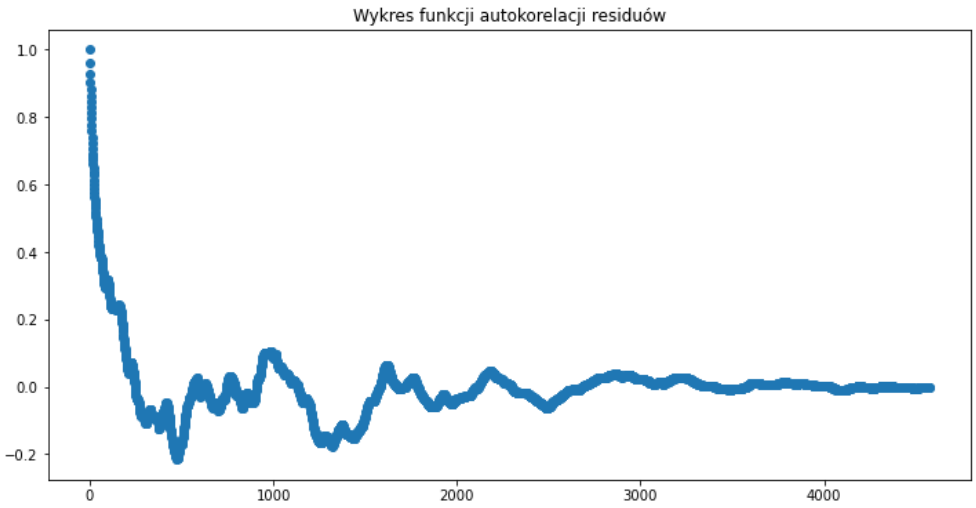
Rysunek 19 – Box-plot wartości residuów.

Box-plot potwierdza, wcześniejszy wniosek o tym, że wartość oczekiwana z residuów, może być stała i równa zero. Dodatkowo, zauważalne są wartości odstające.

4.3 Testowanie Założeń.

1. $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ → niezależne (nieskorelowane);

   Aby sprawdzić niezależność (nieskorelowanie) użyto funkcji autokowariancji i zaprezentowano wyniki na poniższym wykresie (Rys. 20)



Rysunek 20 – Wykres funkcji autokorelacji residuów.

Wykres funkcji autokorelacji pokazuje nam, że wartości tej funkcji są za wysokie aby stwierdzić nieskorelowanie, co także nie zgadza się z jednym z założeń klasycznego modelu regresji liniowe.

2. $\forall_i \, E(\varepsilon_i) = 0$;

   Obliczona średnia residuów wyniosła:

   $$Ee \approx 2.70 \cdot 10^{-13}.$$

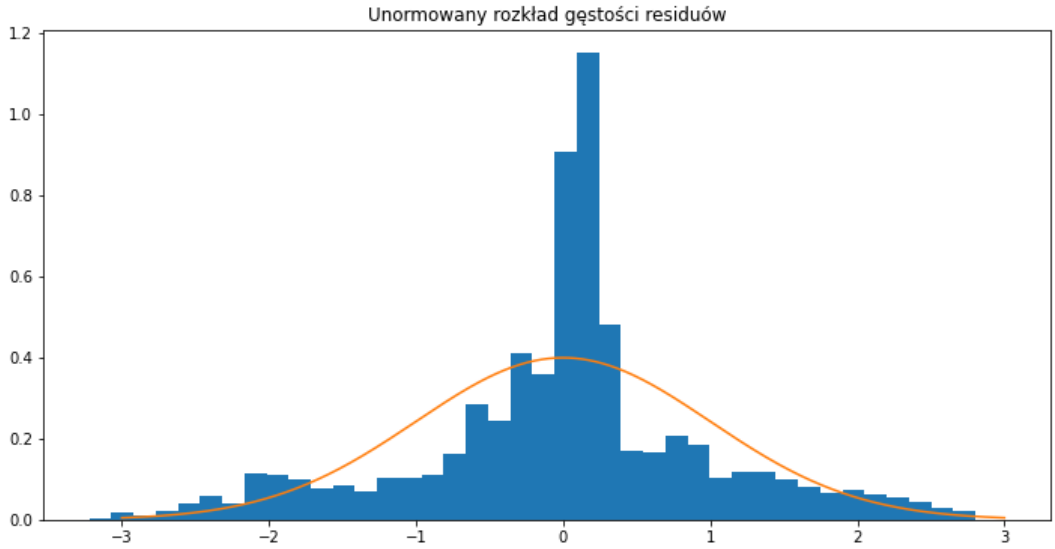
   Taką wartość istotnie możemy przyrównać do 0.

3. $\forall_i \, Var(\varepsilon_i) = \sigma^2$;

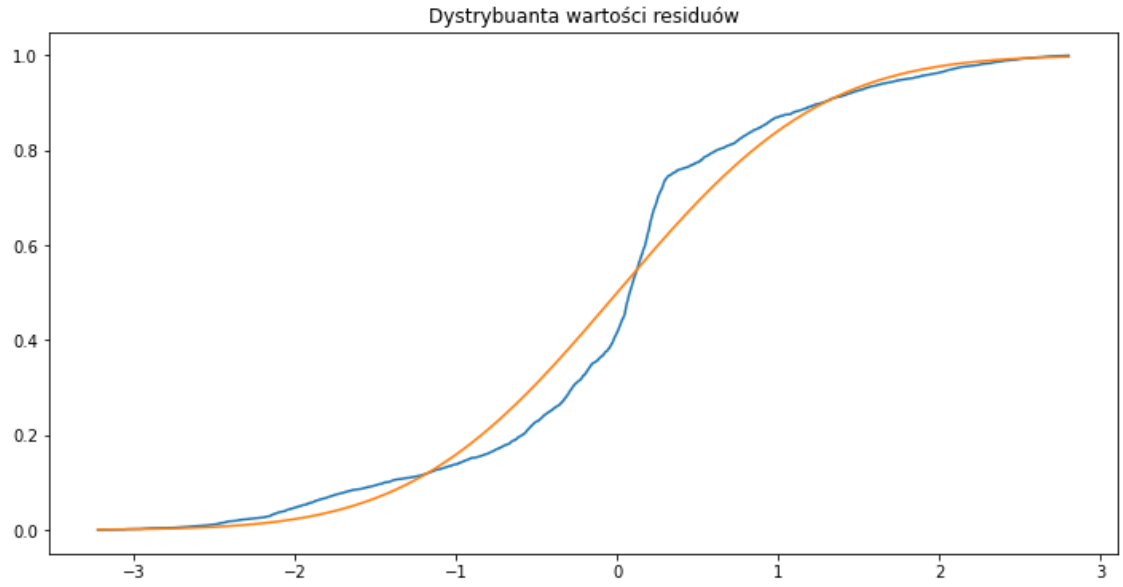
   Aby sprawdzić stałość wariancji skorzystano z testu Bartletta. Otrzymano wynik testu, ze statystyką około 6341.39 i wartością p (p-value) równą zero, wskazuje to na silne dowody przeciwko hipotezie zerowej o równości wariancji w grupach poddanych testowi.

4. $\forall_i \, \varepsilon_i \sim N(0, \sigma^2)$;

   Aby sprawdzić czy rozkłady residuów, są normlane, na początku zgodnie ze wzorem (Rys.17) na normalizacje zmiennej o rozkładzie normlanym, ze średnią μ i wariancją σ^2 , znormalizowano rozkłady residuów, po czym kolejno porównano, gęstości i dystrybuanty, empiryczne i teoretyczne. Wyniki przedstawiono na poniższych wykresach (Rys.21-22)



Rysunek 21 – Unormowany rozkład gęstości residuów.



Rysunek 22 – Dystrybuanta unormowanych wartości residuów.

Z powyższych wykresów (Rys.21-22) możemy wnioskować, że residua nie pochodzą z rozkładu normalnego, ponieważ nie widać znaczącego podobieństwa między miarami teoretycznymi i empirycznymi. Dodatkowo potwierdzają to testy Shapiro-Wilka i Jarque-Bera. Otrzymano wynik testu Shapiro-Wilka, ze statystyką około 0.96 i wartością p (p-value) równą 0, sugeruje to silne dowody przeciw hipotezie zerowej o normalności rozkładu. Natomiast otrzymane wyniki testu Jarque-Bera, ze statystyką 140.67 i wartością p (p-value) równą 0, również wskazują, przeciw hipotezie zerowej o normalności rozkładu.

4.3 Wnioski.

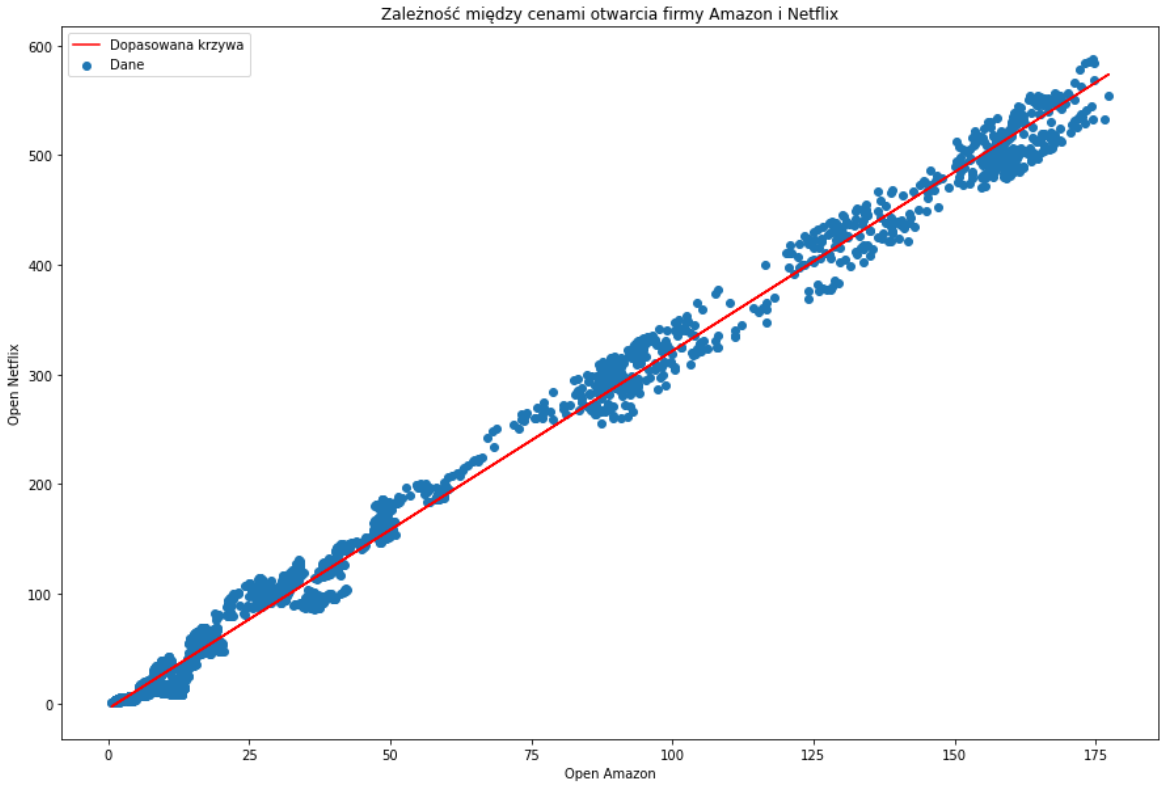
Powyższa analiza residuów, pokazała, że nie pochodzą one z rozkładu normalnego, nie mają stałej wariancji i nie jesteśmy w stanie sprawdzić niezależności (nieskorelowania) danych.

5. Predykcje.

Do przeprowadzenia predykcji wartości przyszłej $Y(\mathrm{x}_0)$ podzielimy nasz zbiór na dwa:

- zbiór treningowy (x_i, y_i) dla $i = 1,\ldots,4522$,
- zbiór testowy (x_i, y_i) dla $i = 4523,\ldots,4572$.

5.1 Dopasowanie nowej prostej.



Rysunek 23- Zbiór treningowy wraz z dopasowaną prostą.

Aby dopasować prostą regresji przedstawioną na powyższym wykresie (Rys.23) skorzystano ze wzorów przedstawionych w rozdziale 3 i za ich pomocą wyliczono współczynniki

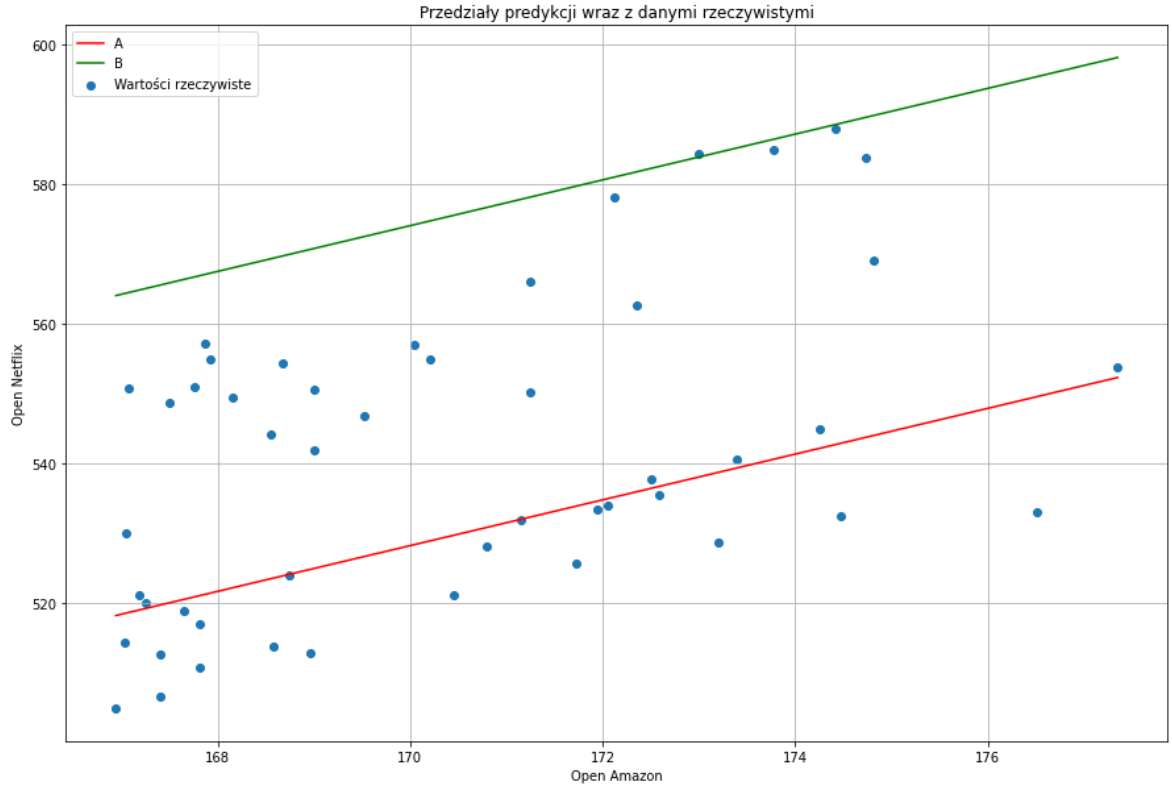
$$\beta_1 = 3.271583621432479 \approx 3.27,$$

$$\beta_0 = -5.003417061744855 \approx -5.00$$

więc naniesiona prosta wyrażona jest wzorem

$$y = 3.27x - 5.0.$$

5.2 Wyznaczenie przedziałów predykcji.



Rysunek 24 - Wyznaczone przedziały z wartościami rzeczywistymi.

Mając wyznaczoną prostą regresji dla zbioru treningowego możemy wyliczyć przedział predykcji dla wartości przyszłej $Y(\mathrm{x}_0)$ na poziomie ufności $1-\alpha$,dla $\alpha = 0.05$ o nieznanej wariancji wyrażony wzorem:

$$\left[\hat{Y}(x_0) - t_{n-2,1-\frac{\alpha}{2}} S \sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, \hat{Y}(x_0) + t_{n-2,1-\frac{\alpha}{2}} S \sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right]$$

gdzie,

- $\hat{Y}(x_0) = \widehat{\beta_1} x_0 + \widehat{\beta_0}$;
- $t_{n-2,1-\frac{\alpha}{2}}$ to kwanty rzędu $1-\frac{\alpha}{2}$, dla α=0.05, z rozkładu t-Studenta z $n-2$ stopniami swobody, a n to długość wektora wartości zbioru treningowego;
- $S^2 = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$.

6. Podsumowanie.

W raporcie analizowaliśmy ceny akcji otwarcia dystrybutora filmów Netflix oraz przedsiębiorstwa Amazon. Naszym zadaniem było dokonanie analizy zależności liniowej tych danych rzeczywistych. Za zmienną niezależną uznaliśmy cenę otwarcia akcji Amazona, za zmienną zależną uznaliśmy cenę akcji Netflix.

Pierwszym krokiem było wykonanie analizy jednowymiarowej. Porównaliśmy gęstości, dystrybuanty i wykresy pudełkowe naszych zmiennych. Jak można zobaczyć na wykresie (Rys.3) czy (Rys.4) wykresy te zachowują podobny charakter na całej długości.

Dalej wyznaczyliśmy wielkości statystyczne dla zmiennych. Ze względu na znaczną różnicę cen zaczynającą się po 2013 roku niektóre parametry bardziej lub mniej różniły się od siebie.

Kolejnym krokiem była analiza zależności liniowej pomiędzy zmienną zależną i zmienną niezależną. Wizualizacja rysunek (Rys.9) pokazała nam zauważalny liniowy rozkład. Przed usunięciem wartości odstających obliczyliśmy współczynniki regresji liniowej, wynoszące $\beta_0 \approx -1.92$ i $\beta_1 \approx 3.20$. Dla nieoczyszczonych danych $R^2 = 0.9463$ co oznacza silny związek korelacji.

Po detekcji i usunięciu obserwacji odstających otrzymaliśmy $\beta_0 = -4.87$ i $\beta_1 = 3.26$. Współczynnik $R^2 = 0.9933$ - wzrósł, co oznacza, że wykryliśmy obserwacje odstające z sukcesem.

Po analizie residuów, pokazano, że nie pochodzą one z rozkładu normalnego, nie mają stałej wariancji i nie jesteśmy w stanie sprawdzić niezależności (nieskorelowania) danych, czyli residua, nie spełniają założeń.

Dokonaliśmy predykcji oraz wyznaczyliśmy przedział ufności dla 50 największych obserwacji. 18 z 50 obserwacji znajdowało się poza przedziałem ufności. Mogło to być

spowodowane wielkością próby lub niespełnionym założeniem, że residua powinny mieć stałą wariancję. Większość danych zawiera się w przedziale ufności.

Po analizie regresji oraz jakości predykcji danych testowych możemy stwierdzić, że nasz model nie spełnia wszystkich założeń go dotyczących. Korzystanie z naszego modelu do predykcji przyszłych wartości firm na giełdzie byłoby ryzykowne. Jednak początkowe dopasowanie prostej świadczy o znacznej możliwości poprawy po ewentualnym usunięciu wartości odstających.