Raport 1

Komputerowa analiza szeregów czasowych

Wykorzystanie poznanych metod służących do analizy zależności liniowej dla wybranych danych rzeczywistych.

Antczak Jakub, Curkowicz Kinga

18.12.2023

Spis treści

1 Opis danych

Niniejszy raport ma na celu przeprowadzenie analizy zależnych od siebie danych dotyczących cen akcji The Walt Disney Company z wykorzystaniem regresji liniowej. Wybrane dane pochodzą ze strony Yahoo Finance, obejmują notowania cen akcji na otwarciu oraz zamknięciu każdego dnia. Naszym głównym celem jest zbadanie zależności między ceną zamknięcia a ceną otwarcia akcji Disney'a. Analiza ta pozwoli nam lepiej zrozumieć, w jaki sposób te dwie zmienne kształtują się w czasie i jakie czynniki mogą wpływać na ich relację.

Badany okres to $18.12.2017 - 18.12.2023$, co zapewniło nam 1508 dni działania firmy Disney. W analizie wykluczyliśmy weekendy oraz święta z tego okresu, gdyż giełda była zamknięta w tych dniach. Ostatnią datą uwzględnioną w naszych danych jest 14 grudnia 2023 roku, ponieważ od 15 do 17 grudnia 2023 roku przypadał weekend, a dane z 18 grudnia 2023 roku zostały pobrane w dniu analizy.

Date	**Open**	High	Low	**Close**	Adj. Close	Volume
2017-12-18	111.839996	111.989998	110.309998	111.029999	107.626457	12270000
2017-12-19	111.330002	112.389999	110.769997	111.809998	108.382553	10546600
2017-12-20	111.620003	112.300003	109.690002	109.690002	106.327538	8661100
2017-12-21	109.720001	111.089996	109.190002	109.570000	106.211212	9372800
2017-12-22	109.489998	109.690002	108.449997	108.669998	105.338799	7378400
2017-12-26	108.489998	109.370003	107.889999	108.120003	104.805672	3982400
2017-12-27	108.419998	108.550003	107.459999	107.639999	104.340370	5624000
2017-12-28	108.000000	108.050003	107.059998	107.769997	104.466393	3477700
2017-12-29	108.050003	108.339996	107.510002	107.510002	104.214363	4538400
2018-01-02	108.949997	111.809998	108.559998	111.800003	108.372864	11014300
2018-01-03	112.190002	113.190002	111.449997	112.279999	108.838135	9237900
2018-01-04	112.949997	113.000000	111.629997	112.230003	108.789688	7417400
2018-01-05	112.680000	112.680000	111.239998	111.620003	108.198380	6008300
2018-01-08	110.889999	111.279999	109.540001	110.019997	106.647423	8052600
2018-01-09	110.129997	110.860001	109.860001	109.940002	106.569878	5838000

Tabela 1: Pobrane dane, Disney - pierwsze 15 obserwacji

Tabela 1 przedstawia udostępniane przez stronę informacje dla każdego dnia działania giełdy, czyli kolejno, cenę otwarcia, najwyższą i najniższą cenę akcji, cenę zamknięcię, skorygowaną cenę zamknięcia, wolumen, czyli ile akcji danej spółki zostało kupionych lub sprzedanych w trakcie sesji giełdowej tego dnia. Wybraną walutą jest dolar amerykański. W dalszej części raportu, skupimy się na dwóch istotnych kolumnach - cenach akcji na otwarciu (**open**) oraz cenach akcji na zamknięciu (**close**).

2 Analiza jednowymiarowa zmiennych

Przejdziemy teraz do analizy jednowymiarowej zmiennej zależnej oraz zmiennej niezależnej. Przyjęliśmy w naszym raporcie cenę akcji Disney'a na otwarciu jako zmienną niezależną oraz cenę akcji Disney'a na zamknięciu jako zmienną zależną. Dokładne wartości liczbowe zostały przedstawione w tabeli 2.

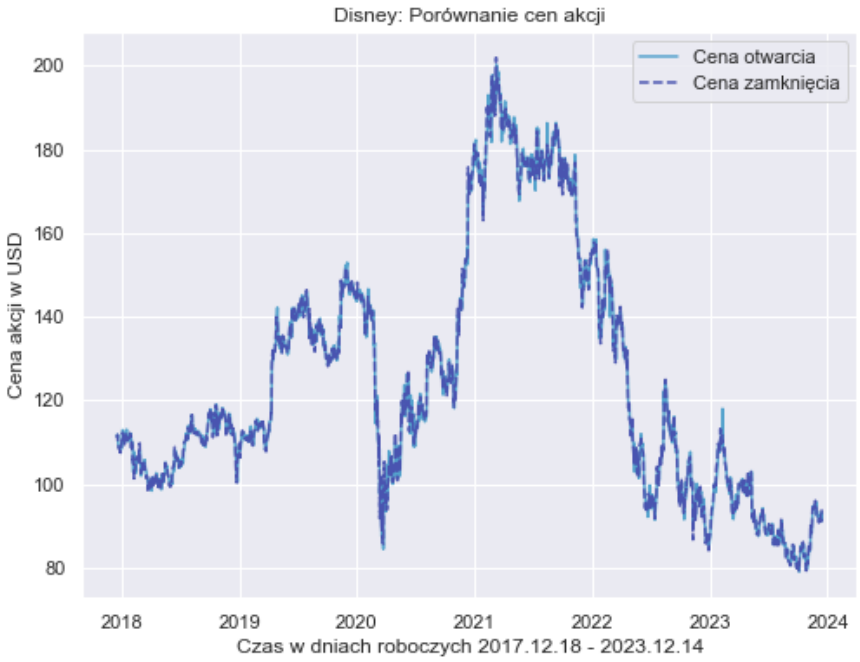
2.1 Wizualizacja

Prezentacja danych została wykonana przy użyciu kilku popularnych bibliotek w języku Python oraz dostępnych w nich funkcji : `NumPy`, `Matplotlib`, `Seaborn` i `SciPy`.

2.1.1 Porównanie cen akcji

Na początku przedstawiliśmy na Wykresie 1, jak zmieniały się ceny otwarcia i zamknięcia firmy Disney w badanym okresie, czyli od 18 grudnia 2017 roku do 14 grudnia 2023 roku. Możemy zauważyć, że różnica

pomiędzy ceną otwarcia i ceną zamknięcia w poszczególnych dniach jest bardzo mała. Było to zgodne z naszą intuicją, ponieważ analizowane zmienne są dużym stopniu od siebie zależne i wzajemnie proporcjonalne.



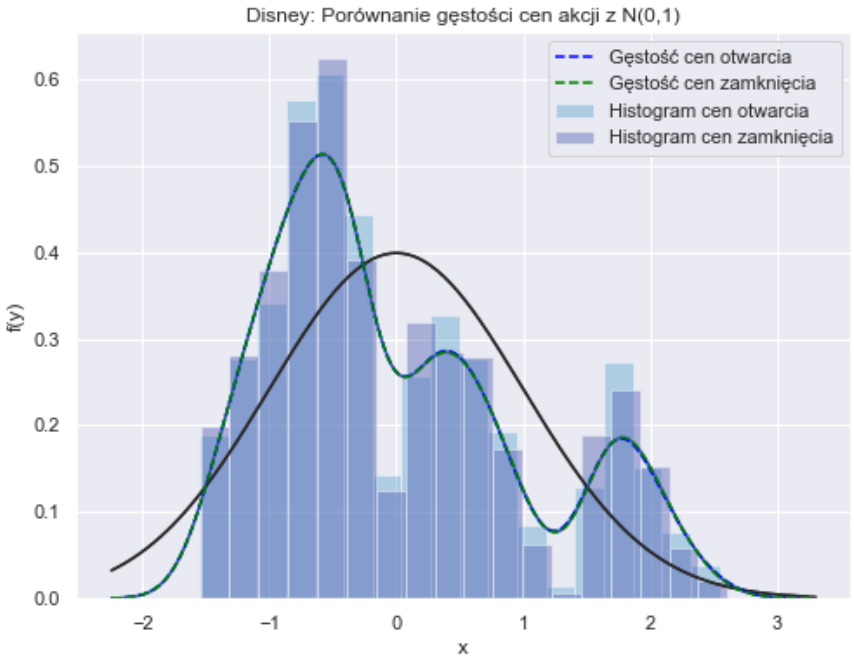
Wykres 1: Ceny akcji Disney'a na otwarcie i zamknięcie

2.1.2 Porównanie gęstości

Następnie porównaliśmy gęstości cen otwarcia oraz cen zamknięcia akcji. Przedstawione jest to na Wykresie 2. Ustandaryzowaliśmy nasze dane, żebyśmy mogli porównać rozkłady między sobą oraz z rozkładem normalnym z parametrami $\mu = 0$ i $\sigma = 1$, gdzie μ to średnia, a σ odchylenie standardowe. Średnia z danych w przypadku cen otwarcia wynosiła 124.97, a odchylenie standardowe było równe w przybliżeniu 29.57. Dla cen zamknięcia otrzymaliśmy średnią 124.87 oraz odchylenie standardowe 29.51, wykorzystaliśmy te wartości przy standaryzacji.

Rozkład gęstości cen otwarcia praktycznie na całej swej długości jest zbliżony do rozkładu gęstości cen zamknięcia. To zgodne z intuicją, gdyż ceny akcji w chwili otwarcia i zamknięcia rynku podlegają wpływom gospodarki rynkowej w sposób bardzo zbliżony. W związku z tym ich rozkłady gęstości wykazują podobne charakterystyki.

Gęstości analizowanych danych znacząco odbiegają od typowego symetrycznego kształtu krzywej Gaussa, co świadczy o bardziej złożonych wzorcach wpływów gospodarki na rynku akcji. W rezultacie możemy wnioskować, że dane te nie są generowane z rozkładów zbliżonych do normalnego, gdyż wykresy znacząco się różnią.



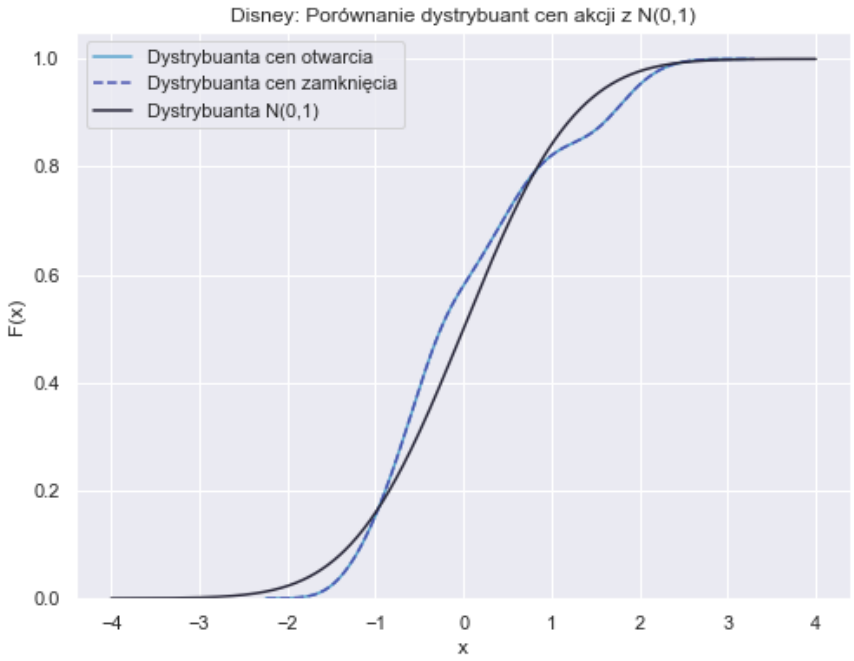
Wykres 2: Ceny akcji Disney'a na otwarcie i zamknięcie

2.1.3 Porównanie dystrybuant

Na Wykresie 3 przedstawiono empiryczne dystrybuanty cen otwarcia i zamknięcia akcji, które zostały poddane standaryzacji w celu porównania z dystrybuantą rozkładu normalnego o średniej $\mu = 0$ i odchyleniu standardowym $\sigma = 1$. Standaryzacja jednakowa jak w przypadku porównania gęstości.

Zauważamy, że dystrybuanty empiryczne cen otwarcia i zamknięcia znacznie się nakładają, co jest zgodne z wcześniejszą analizą gęstości danych oraz obserwacją zmienności w czasie. Z tego powodu zostały przedstawione za pomocą przerywanej lini.

Jednakże, ważne jest spostrzerzenie, że żadne z odcinków dystrybuant empirycznych nie pokrywają się z dystrybuantą rozkładu normalnego. Ponownie sugeruje nam to, że dane te prawdopodobnie nie są generowane z rozkładu zbliżonego do normalnego.



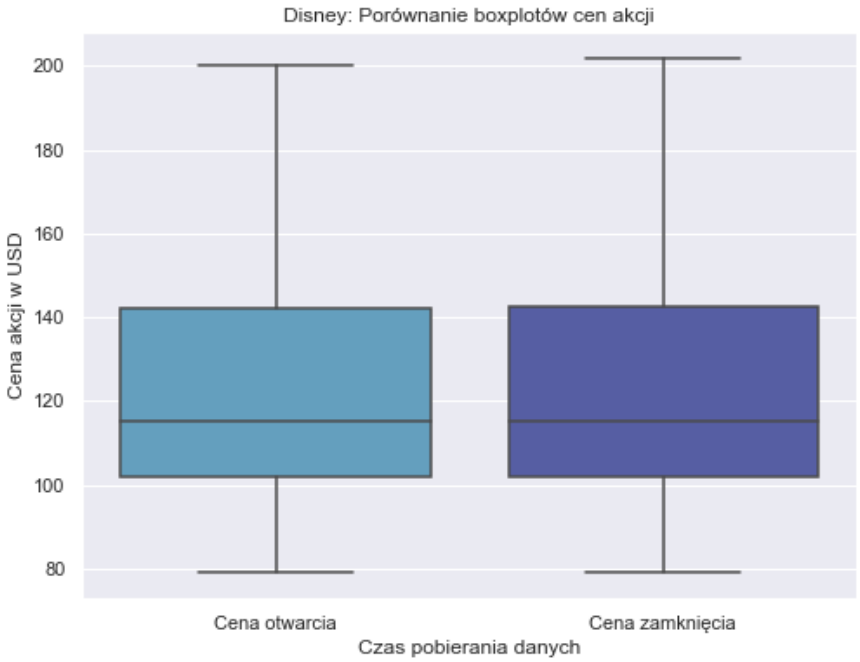
Wykres 3: Ceny akcji Disney'a na otwarcie i zamknięcie

2.1.4 Porównanie wykresów pudełkowych

Na Wykresie 4 prezentujemy porównanie wykresów pudełkowych cen Disney'a na otwarciu akcji, oraz na ich zamknięciu. Na pierwszy rzut oka, nie widać znacznych różnic pomiędzy wykresami, co jest zgodne z intuicją, gdyż zmiennna zależna i niezależne są do siebie bardzo zbliżone. Patrząc na przedstawiony wykres możemy stwierdzić, że mediany w obu przypadkach są bardzo podobne, niemal identyczne. Nie zaobserwowaliśmy, żadnych obserwacji odstających.

Długie wąsy wskazują, że wartości w obu przypadkach są rozproszone i zajmują zakres od około 80 do około 200. Ten zakres danych pokrywa się z Wykresem 1 przedstawiającym porównanie cen akcji przez sześcioletni okres naszej analizy.

Dodatkowo, rozstęp międzykwartylowy, czyli różnica między trzecim a pierwszym kwartylem, również jest zbliżony dla cen otwarcia i zamknięcia. Obejmuje on zakres cen od około 100 do 140 USD. Istotne jest, że nie ma żadnych wartości odstających, co sugeruje, że dane są dość stabilne i nie występują znaczne anomalie w cenach akcji.



Wykres 4: Ceny akcji Disney'a na otwarcie i zamknięcie

2.2 Statystyki opisowe

Istotnym aspektem analizy rzeczywistych danych są statystyki opisowe, które charakteryzują właściwości badanej próby. Stanowią one narzędzie do skondensowanego podsumowania informacji, umożliwiającego prawidłowe wnioskowanie. Przedstawimy podstawowe miary położenia, rozproszenia, skośności, spłaszczenia oraz odpowiadające im wzory.

2.2.1 Miary położenia

Miary położenia wskazują wokół jakich wartości oscylują analizowane dane. Wyróżnia się miary klasyczne, takie jak średnie oraz pozycyjne, czyli na przykład kwartyle. Obie te miary opisują pewne własności i cechy próby z innego punktu widzenia.

Dwie spośród znanych średnich to średnia arytmetyczna

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

oraz średnia ucinana

$$\bar{x}_u = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} x_i. \tag{2}$$

Kwartyle natomiast to miary, które dzielą zbiór danych na cztery grupy. Wyróżniamy następujące kwartyle:

- drugi kwartyl $Q2$
  Jest to szczególny kwartyl, nazywany inaczej medianą, który dzieli zbiór na dwa równoliczne podzbiory

i wyraża się następującym wzorem

$$x_{med} = \begin{cases} x_{((n+1)/2)}, & \text{n nieparzyste} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}), & \text{n parzyste.} \end{cases} \tag{3}$$

- pierwszy kwartyl $Q1$ (mediana grupy obserwacji mniejszych od Q2)
- trzeci kwartyl $Q3$ (mediana grupy obserwacji większych od Q2)

2.2.2 Miary rozproszenia

Miary rozproszenia opisują, jak zróżnicowane są wartości w danej próbie. Zaliczają się do nich:

- rozstęp międzykwartylowy
$$IQR = Q3 - Q1 \tag{4}$$
- rozstęp z próby
$$R = x_{(n)} - x_{(1)} \tag{5}$$
- wariancja
$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{6}$$
- odchylenie standardowe
$$S = \sqrt{S^2} \tag{7}$$
- współczynnik zmienności
$$V = \frac{S}{\bar{x}} (\cdot 100\%) \tag{8}$$

2.2.3 Miary skośności i spłaszczenia

- współczynnik skośności - miara asymetrii
$$\alpha = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S} \right)^3 \tag{9}$$
- kurtoza - miara spłaszczenia
$$K = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^2} \tag{10}$$

2.2.4 Wielkości statystyczne i ich interpretacja

Wartości poszczególnych statystyk opisowych dla ceny otwarcia oraz zamknięcia przedstawiono poniżej w tabeli 2.

Podsumowując analizę statystyk opisowych cen akcji firmy Disney można wyróżnić kilka kluczowych obserwacji. Średnie ceny otwarcia i zamknięcia utrzymywały się na poziomie około 124.97 i 124.87, odpowiednio, co tak jak większość parametrów dla ceny otwarcia i zamknięcia różni się między sobą o maksymalnie wielkość rzędu części dziesiętnych, co potwierdza ich zależność.

Wyznaczone mediany dla obu zmiennych są mniejsze od ich średnich arytmetycznych, co wskazuje na prawoskośność danych, o czym świadczy również dodatnia wartość współczynnika skośności. Naszą obserwację możemy zauważyć również na Wykresie 2.

Wartości odchylenia standardowego wskazują na umiarkowaną zmienność cen, co jest zgodne z charakterem rynków finansowych, kiedy nie mamy doczynienia z krachem giełdowym. Odchylenie standardowe w

	Otwarcie	Zamknięcie
Średnia arytmetyczna	124,97	124,87
Odchylenie standardowe	29,57	29,51
Minimum	79,10	79,32
25. percentyl	101,99	102,11
Mediana	115,19	115,21
75. percentyl	142,28	142,51
Maksimum	200,19	201,91
Skośność	0,70	0,70
Kurtoza	-0,50	-0,51
Współczynnik zmienności	23,66%	23,63%

Tabela 2: Statystyki opisowe dla cen otwarcia i zamknięcia Disney'a

porównaniu do średniej jest stosunkowo małe, zatem możemy uznać, że ceny akcji firmy Disney w analizowanym okresie wykazują umiarkowaną zmienność wokół średniej wartości, co może świadczyć o pewnej stabilności cenowej.

Ze względu na ujemną kurtozę w obu przypadkach można stwierdzić, że rozkłady cen otwarcia i zamknięcia akcji Disney'a są bardziej spłaszczone niż w przypadku rozkładu normalnego. Kurtoza mierzy stopień spłaszczenia rozkładu w porównaniu do rozkładu normalnego, który ma kurtozę równą 0.

2.3 Interpretacja wyników

Podsumowanie analizy danych pozwala nam wyciągnąć kilka istotnych wniosków. Przede wszystkim, zauważamy niemal identyczne charakterystyki danych w każdym przypadku, co odzwierciedlają zbieżne wykresy i równoważne statystyki. Wnioskiem z tego jest przypuszczenie, że analizowane dane pochodzą z tego samego rozkładu. Możemy stwierdzić, że rozkład cen otwarcia akcji oraz rozkład cen zamknięcia Disney'a nie jest rozkładem normalnym oraz dane nie zachowują symetrii.

Pierwsza obserwacja wynika z faktu, że gęstość cen otwarcia nie jest symetryczna i wykazuje pewne nieregularności, co jest charakterystyczne dla rozkładów, które nie są normalne. Dodatkowo, analiza gęstości cen zamknięcia potwierdza tę obserwację.

Druga obserwacja dotyczy nieregularnego kształtu gęstości cen zamknięcia, co jest także sprzeczne z założeniem o normalności rozkładu. Oba rozkłady zdają się mieć dwie wyraźne "szczytowe" części, co sugeruje, że rozkład może być bimodalny lub złożony z co najmniej dwóch składowych.

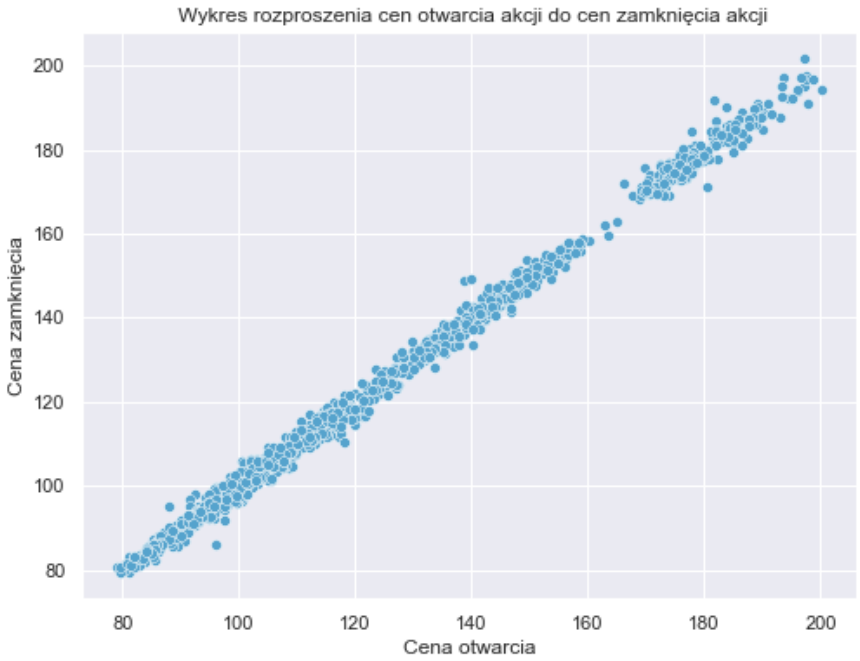
3 Analiza zależności liniowej pomiędzy zmienną zależną a zmienną niezależną

Przypomnijmy, że przyjęliśmy cenę otwarcia akcji Disney'a jest jako zmienną niezależną, natomiast cenę zamknięcia jako zmienną zależna. Zależy nam na wyznaczeniu ceny zamknięcia od ceny otwarcia.

3.1 Prezentacja danych

Przedstawimy teraz nasze dane na Wykresie 5, czyli wizualizacji rozproszenia cen otwarcia akcji do cen zamknięcia akcji Disney'a.

Obserwujemy niemalże idealną zależność liniową pomiędzy zmiennymi. Jest to zgodne z naszymi oczekiwaniami, ponieważ obie wielkości w rzeczywistości często różnią się o niewielką wartość, oczywiście gdy nie mamy do czynienia z krachem giełdowym.



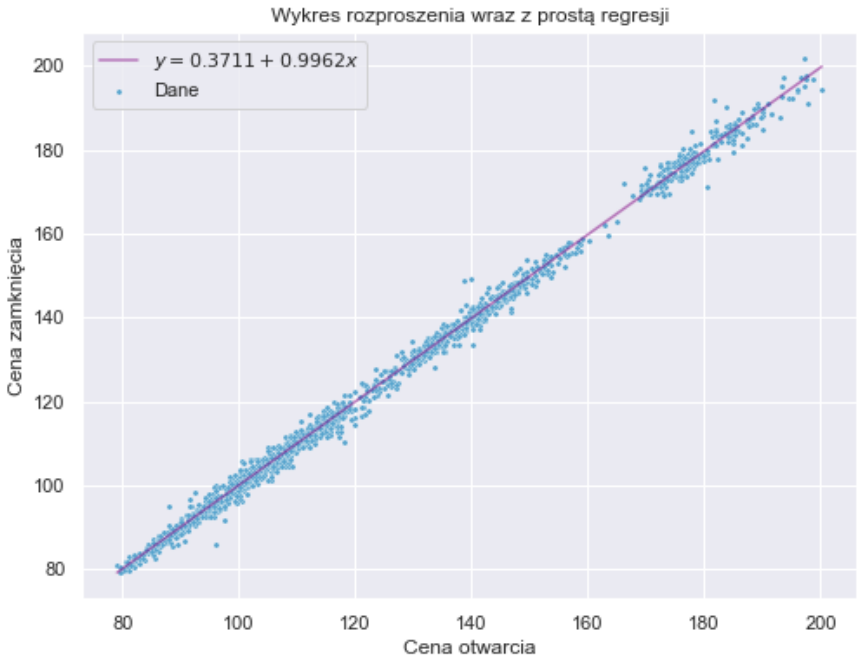
Wykres 5: Wykres rozproszenia danych

3.2 Prosta regresji

Aby wyznaczyć prostą regresji, posłużyliśmy się metodą najmniejszych kwadratów do wyznaczenia współczynników prostej. Założyliśmy, że prosta ma postać $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, szukamy takich $\hat{\beta}_0, \hat{\beta}_1$, które minimalizują sumę kwadratów błędów $S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$, gdzie y_i to kolejne ceny zamknięcia akcji, a x_i otwarcia akcji. Po obliczeniu pochodnych po obu współczynnikach $\hat{\beta}_0$, $\hat{\beta}_1$, przyrównaniu ich do zera oraz dokonaniu kilku przekształceń otrzymaliśmy następujące wzory oraz wyniki.

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \approx 0.3711, \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x} \cdot \overline{y}}{\sum_{i=1}^{n} x_i^2 - n(\overline{x})^2} \approx 0.9962. \tag{11}$$

Analizowane dane wraz z wyznaczoną prostą regresji $\hat{y} = 0.3711 + 0.9962x$ przedstawiono na Wykresie 6.



Wykres 6: Wykres rozproszenia wraz z dopasowaną prostą regresji

Wartości współczynników zgadzają się z intuicją. Cena otwarcia i cena zamknięcia nie różnią się znacząco, a w dodatku po zobrazowaniu danych na Wykresie 6, łatwo zauważyć, że punkty leżą blisko prostej $y = x$, co potwierdza ich względną bliskość wartości pomiędzy cenami otwarcia i zamknięcia akcji. Wartości współczynników regresji liniowej $\hat{\beta}_0$ i $\hat{\beta}_1$ są zgodne z tą obserwacją. W przypadku, gdy $\hat{\beta}_0$ jest bliskie zeru, a $\hat{\beta}_1$ zbliżone do jedynki to możemy wnioskować, że istnieje silna liniowa zależność między ceną otwarcia a zamknięcia.

Wyznaczyliśmy również wartość $R^2 = 0.996$ opisaną wzorem (21), którą wykorzystamy do porównania w podsumowaniu 5.

Oczekujemy, że po usunięciu wartości odstających współczynniki prostej regresji jeszcze dokładniej odzwierciedlą tę zależność. Sprawdzimy to w kolejnym punkcie raportu 3.3.

3.3 Detekcja obserwacji odstających

Przejdziemy teraz do analizy danych bez uwzględniania wartości odstających, ponieważ dni wprowadzenia usługi Disney+ mogą charakteryzować się wyjątkowo wysoką zmiennością cen akcji. Jest to rezultat intensywnego zainteresowania rynkowego nową usługą, co potencjalnie wpływa na dynamiczne ruchy cenowe. Poprzez wyłączenie wartości odstających, dążymy do lepszego zrozumienia trwalszych trendów cenowych, pomijając skrajne wahania związane z wprowadzeniem Disney+ na rynek.

Wprowadzenie usługi Disney+ na rynek miało miejsce w listopadzie 2019 roku w Stanach Zjednoczonych, a następnie stopniowo ekspandowano ją do innych krajów na świecie. Platforma dotarła do największej ilości nowych krajów na przelomie roku 2021/2022, co obrazuje Wykres 1 oraz tabela 3 zawierająca 10 przykładowych obserwacji odstających z 59 znalezionych.

Nr	Data	Cena Otwarcia	Cena Zamknięcia
752	2020-12-14	173.800003	169.300003
765	2021-01-04	182.259995	177.679993
782	2021-01-28	166.169998	171.880005
789	2021-02-08	183.850006	190.000000
793	2021-02-12	193.000000	187.669998
798	2021-02-22	181.740005	191.759995
801	2021-02-25	197.729996	190.979996
808	2021-03-08	197.309998	201.910004
809	2021-03-09	200.190002	194.509995
820	2021-03-24	190.059998	184.720001

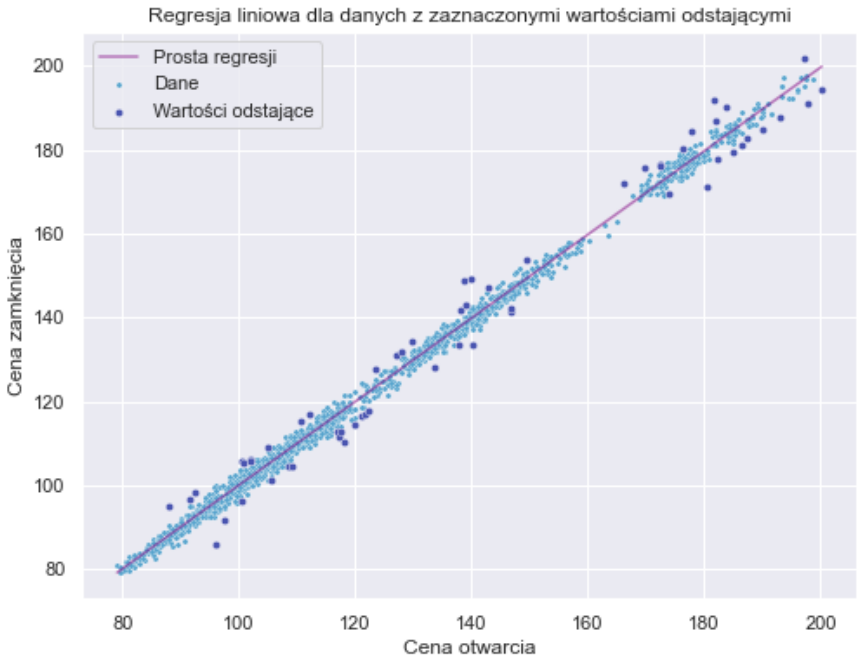
Tabela 3: Dane dotyczące cen akcji Disney w określonych dniach.

Zdefiniujmy zatem residua, które wykorzystamy do dalszej analizy.
Residua modelu regresji liniowej definiujemy jako ciąg $\{\epsilon_i\}_{i=1}^{n}$, gdzie:

$$\epsilon_i = Y_i - \hat{Y}_i, i = 1, \ldots, n \tag{12}$$

Dla tak wyznaczonych residuów będziemy badać, czy występują dla nich obserwacje odstające. W przypadku modelu regresji liniowej, wartości odstające odpowiadają residuom, których wartości bezwzględne przekraczają 1.5-krotność rozstępu międzykwartylowego danego wzorem (4).

Zatem dla resiudów naszego modelu uznajemy wartości odstające, które wykraczają poza następujący przedział $[Q1-1.5IQR, Q3+1.5IQR]$, gdzie $Q1$ to pierwszy kwartyl, $Q3$ to trzeci kwartyl, a IQR to rozstęp międzykwartylowy.

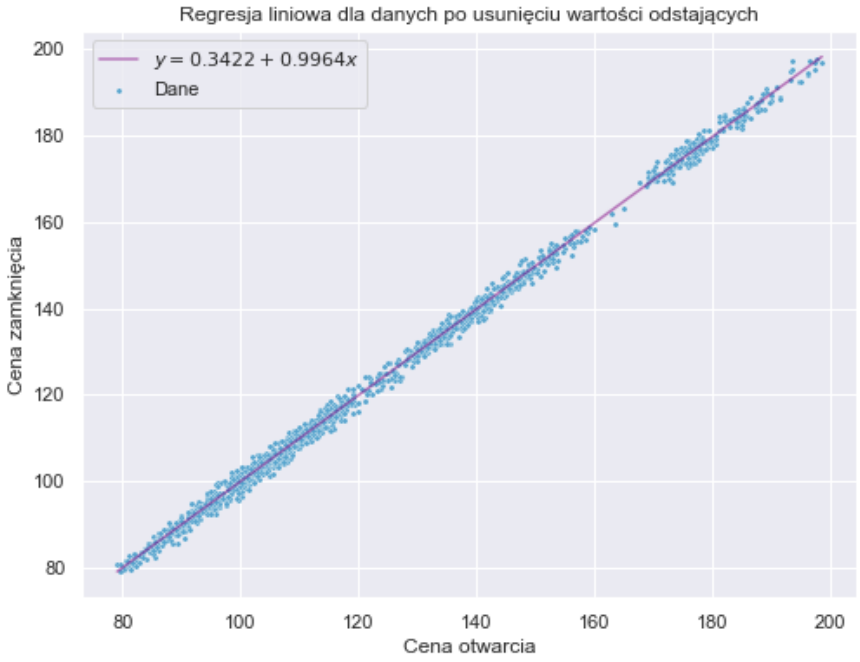


Wykres 7: Wykres rozproszenia wraz z dopasowaną prostą regresji i obserwacjami odstającymi

3.4 Wyznaczenie nowych współczynników prostej regresji

Po zastosowaniu metody detekcji wartości odstających opisanej w 3.3 otrzymaliśmy 59 wartości odstających. Usuneliśmy je i ponownie obliczyliśmy współczynniki prostej regresji za pomocę metody najmniejszych kwadratów. Dla wzoru funkcji liniowej $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, otrzymujemy następujące wartości, $\hat{\beta}_0 = 0.3422$, $\hat{\beta}_1 = 0.9964$

To właśnie tych wartości współczynników będziemy używać do obliczeń w dalszej części raportu. Zgodnie z oczekiwaniami, po usunięciu wartości odstających nasze współczynniki jeszcze bardziej przybliżyły wzór prostej regresji do postaci $y = x$, ponieważ wartość $\hat{\beta}_0$ jest w zaokrągleniu do tych samych miejsc po przecinku bliżej zera, a wartość $\hat{\beta}_1$ bliżej jedynki. Wykres 8 przedstawia prostą regresji wykonaną dla danych po usunięciu obserwacji odstających.



Wykres 8: Wykres rozproszenia wraz z dopasowaną prostą regresji dla oczyszczonych danych

3.5 Wyznaczanie przedziału ufności dla współczynnika β_1

Przedział ufności to przedział, w którym z określonym prawdopodbieństwem, zawiera się prawdziwa wartość pewnego parametru. Chcemy wyznaczyć przedział ufności dla parametru β_1. Weźmy statytykę T określoną wzorem:

$$T = \frac{\hat{\beta}_1 - \beta_1}{S} \sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{13}$$

T ma rozkład t-Studenta z $n-2$ stopniami swobody. Parametr n to długość próby, czyli w tym przypadku jest to $n = 1449$. We wzorze S^2 to estymator wariancji dany wzorem (6).

Chcemy znaleźć przedział ufności, taki że:

$$P(A \leq \beta_1 \leq B) = 1 - \alpha \tag{14}$$

Weźmy $\alpha = 0.05$. W związku z tym, że T pochodzi z rozkładu t-Studenta z $n-2$ stopniami swobody, wiemy, że:

$$P(-t_{n-2,1-\frac{\alpha}{2}} \leq \beta_1 \leq t_{n-2,1-\frac{\alpha}{2}}) \tag{15}$$

Po podstawowych przekształceniach, wartość A wynosi 0.993, a wartość B wynosi 0.999. Ostatecznie $\beta_1 \in [0.993, 0.999]$.

3.6 Wyznaczanie przedziału ufności dla współczynnika β_0

Dla parametru β_0 statystyka testowa T określona jest wzorem:

$$T = \frac{\hat{\beta}_0 - \beta_0}{S} \left(\frac{1}{n} + \frac{(\overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)^{-\frac{1}{2}}. \tag{16}$$

Wykonując te same przekształcenia, co w powyższym punkcie, ustalono, że $\beta_0 \in [0.0115, 0.6727]$

3.7 Ocena poziomu zależności

W celu oceny naszego modelu wyznaczymy podstawowe metryki:

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}} \tag{17}$$

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2 \tag{18}$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \tag{19}$$

$$SSR = \sum_{i=1}^{n} (\hat{y_i}^2 - \overline{y}) \tag{20}$$

$$R^2 = \frac{SSR}{SST} \tag{21}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y_i})^2} \tag{22}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y_i}| \tag{23}$$

Metryka	**Wartość**
Wsp. korelacji Pearsona	0.998
SST	1229513.046
SSE	3112.615
SSR	1226400.430
R^2	0.997
RMSE	1.465
MAE	1.149

Tabela 4: Wyznaczone wartości metryk

Współczynnik korelacji Pearsona r oraz R^2 są prawie równe 1, co nie jest zaskoczeniem. Dane są zależne liniowe i odchylenia od normy są rzadko obserwowane. Można zauważyć również, że SST jest prawie równe $SSR + SSE$, zatem można wnioskować, że wyniki są poprawne.

Informacje, które są dla nas cenne, to z pewnością RMSE i MAE. RMSE, czyli pierwiastek błędu średniokwadratowego mówi o tym, że nasza predykcja różni się od rzeczywistej wartości ceny zamknięcia o 1.47 dolara. Jest to stosunkowo dobry wynik, biorąc pod uwagę wysokość cen. Wartość średniego błędu bezwzględnego, czyli MAE, wynosi około 1.15 dolara i ma o tyle ważne znaczenie, że nie dokonujemy na błędach transformacji typu podniesienie do kwadratu czy pierwiastkowanie. Należy pamiętać również, że wartości odstające, które mogłyby zaburzać wyniki, zostały usunięte.

3.8 Predykcja oraz przedziały ufności dla pewnej części danych

Aby przeprowadzić predykcję danych, wyselekcjonowaliśmy $m = 50$ największych obserwacji cen otwarcia, a następnie na podstawie pozostałych danych dopasowaliśmy prostą regresji.

Niech x_0 będzie naszymi 50 danymi, chcąc wyznaczyć przedział ufności, skorzystamy z faktu:

$$E[Y(x_0) - \hat{Y}(x_0)] = 0 \tag{24}$$

$$Var[Y(x_0) - \hat{Y}(x_0)] = \sigma^2(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=0}^{n}(x_i - \overline{x})^2}) \tag{25}$$

Umiemy teraz skostruować naszą statystykę testową T.

$$T = \frac{(Y(x_0) - \hat{Y}(x_0)) - 0}{S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=0}^{n}(x_i - \overline{x})^2}}} \tag{26}$$

Wyznaczając przedział ufności dla np. β_1, działaliśmy tylko na jednej wartości. W tym przypadku mamy do czynienia z 50 wartościami, dlatego otrzymane wyniki, nie będą już konkretnym przedziałem, a polem znajdującym się między dwoma prostymi. Proste te, muszą spełniać warunek:

$$P(\hat{Y}(x_0) - t_{1-\alpha,n-2}S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=0}^{n}(x_i - \overline{x})^2}} \leq Y(x_0) \leq \tag{27}$$

$$\leq \hat{Y}(x_0) + t_{1-\alpha,n-2}S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=0}^{n}(x_i - \overline{x})^2}}) = 1 - \alpha$$

Weźmy $\alpha = 0.05$. Po obliczeniu potrzebnych wartości, otrzymujemy Wykres 9. Widać, że niektóre z obserwacji znajdują się poza tym przedziałem, bądź leżą idealnie na prostych. Może być to spowodowane tym, że nasze residua nie pochodzą z rozkładu normalnego. Jeśli, nie są, to oznacza, że nasza statystyka nie ma rozkładu t-Studenta, przez co, przedziały ufności mogą być niepoprawne. Jednak nie zmienia to faktu, że większość obserwacji leży wsród przedziału ufności, co może oznaczać, że nasz model jest stabilny na usuwanie ze zbioru części danych.

Wykres 9: Predykcja odłożonej próbki wielkości m, z zaznaczonymi przedziałami ufności na poziomie ufności 1-α

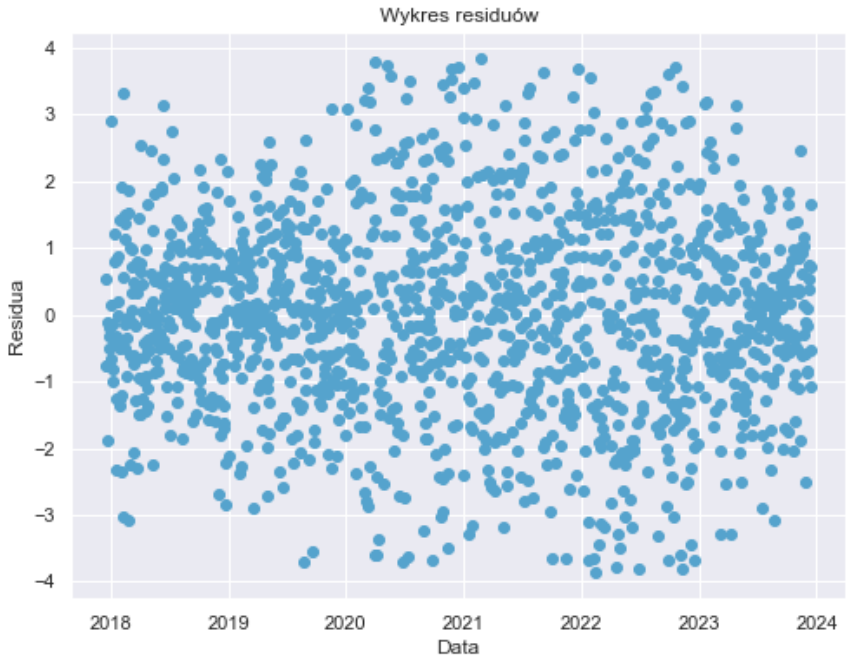
4 Analiza residuów

Definicję residuów modelu możemy znaleźć w sekcji 3.3. W tym punkcie skupimy się na analizie residuów. Założenia, które muszą spełniać residua, to:

1. Czy ich wartość oczekiwana wynosi 0?
2. Czy mają stałą wariancję?
3. Czy są nieskorelowane?
4. Czy pochodzą z rozkładu normalnego?

4.1 Wartość średnia

Zacznijmy od sprawdzenia, czy wartość średnia residuów jest równa 0. Wykres residuów, obliczonych ze wzoru 12, jest widoczny na Wykresie 10. Widać, że residua rozkładają się równomiernie względem zera.

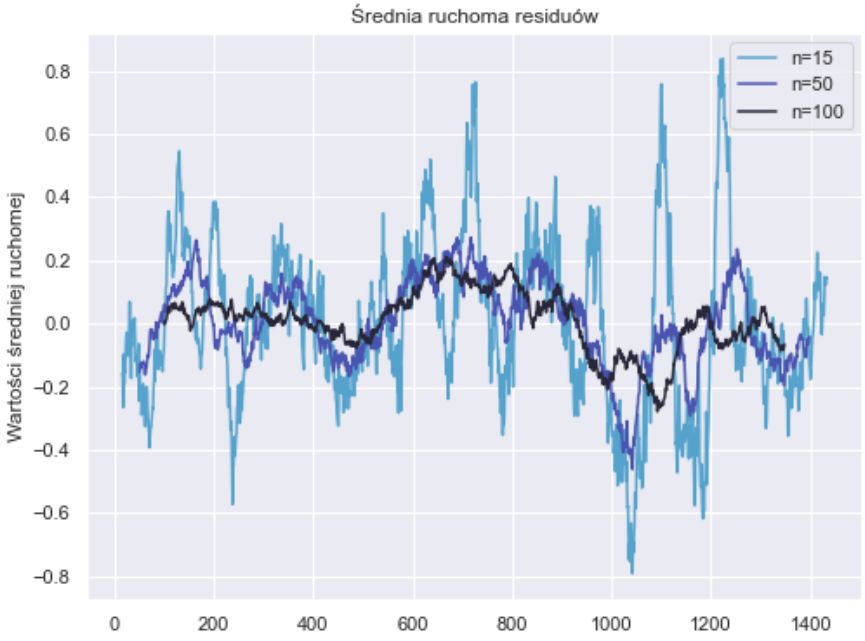


Wykres 10: Wykres residuów

Wyliczona wartość średnia z residuów wynosi $2.266 \cdot 10^{-13}$. Jest to wartość bardzo bliska zeru. Narysowano również wykres średniej ruchomej Wykres 11, która liczy średnią arytmetyczną dla n ostatnich okresów. Aby go narysować, najpierw obliczyliśmy średnią ruchomą zgodnie ze wzorem:

$$MA = \frac{1}{2p+1} \sum_{k=i-p}^{i+p} x_k, \tag{28}$$

gdzie p to n ostatnich okresów. W naszym przypadku $n = 15, 50, 100$.



Wykres 11: Średnia ruchoma residuów

Na wykresie widać, że średnia arytmetyczna z n ostatnich obserwacji, dla każdego z analizowanych dni, układa się prawie symetrycznie względem zera. W dodatku wraz ze wzrostem długości n, wykres wypłaszcza się. Zgodnie z poprzednimi obliczeniami, jeśli za n przyjmiemy długość bliską długości wektora residuów, to otrzmamy liczbę bliską 0, co prowadzi nas do postawienia hipotezy, że residua mają wartość średnią 0.

W celu potwierdzenia, że wartość średnia wynosi 0, przeprowadźmy test t-Studenta na poziomie istotności $\alpha = 0.05$. Weźmy:

$$H_0 : \text{wartość średnia residuów jest równa 0,}$$

$$H_1 : \text{wartość średnia residuów nie jest równa 0.}$$

Do przeprowadzenia testu użyjemy funkcji `ttest_1samp` z biblioteki `scipy.stats` w Pythonie. W ten sposób uzyskaliśmy p-wartość równą około 0.999. Jest to znacznie więcej niż ustalone $\alpha = 0.05$. Funkcja zwraca również t-statystykę, czyli stosunek odchylenia szacowanej wartości parametru od jego hipotetycznej wartości do jego błędu standardowego. W tym przypadku jest ona równa $5.883 \cdot 10^{-12}$. Otrzymane wyniki wskazują na to, że nie ma najmniejszych podstaw do odrzucenia hipotezy zerowej.

4.2 Wariancja

W kolejnym kroku przeprowadzanej analizy, sprawdzimy, czy badane residua mają stałą wariancję. Już po samym wykresie residuów Wykres 10 można stwierdzić, że tak nie jest. Przeprowadźmy test statystyczny Breusha-Pagana, który służy do testowania heteroskedastyczności. Ustalmy poziom istotności $\alpha = 0.05$ oraz przyjmijmy:

$$H_0 : \text{wariancja jest stała,}$$

$$H_1 : \text{wariancja nie jest stała.}$$

W celu przeprowadzenia testu użyjemy funkcji `het_breuschpagan` z biblioteki `statsmodels` w Pythonie. Otrzymaliśmy p-wartość równą $4.555 \cdot 10^{-205}$. Jest to znacznie mniejsze od poziomu istotności α, dlatego też jesteśmy zmuszeni odrzucić hipotezę zerową i przyjąć hipotezę alternatywną.

4.3 Autokorelacja

Kolejnym krokiem analizy jest sprawdzenie, czy residua są od siebie niezależne. Użyjemy funkcji `plot_acf` z biblioteki `statsmodels` w Pythonie, aby narysować funkcję autokorelacji dla residuów. Zdefiniujmy funkcję autokorelacji.

Funkcja autokorelacji (ρ) przy przesunięciu h, gdzie h przyjmuje wartości od $-n$ do n (gdzie n to liczba residuów), mierzy stopień korelacji między wartościami residuów oddalonymi o h przedziałów od siebie w szeregach czasowych. W przypadku teoretycznej funkcji autokorelacji (ρ), dla $h = 0$, wartość wynosi 1, co oznacza doskonałą korelację z samym sobą. Dla $h \neq 0$, wartość wynosi 0, co oznacza brak korelacji.

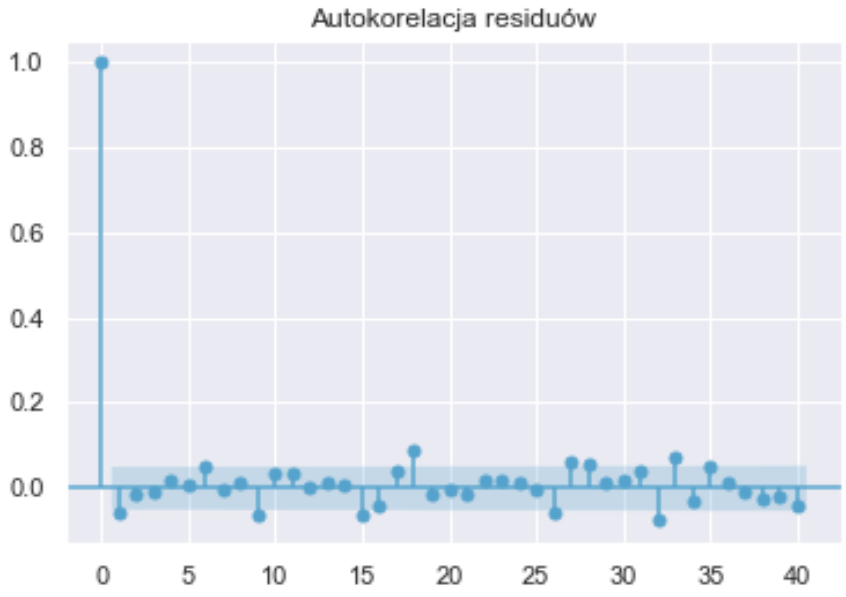
Będziemy porównywać ją z jej empirycznym odpowiednikiem:

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} \quad (21)$$

gdzie $\hat{\gamma}$ to empiryczna funkcja autokowariancji dana wzorem:

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{i=1}^{n-|h|} \left(e_{i+|h|} - \bar{e}\right)(e_i - \bar{e}) \quad (22)$$

Weźmy wektor $h = 0 : 1 : 40$. Sprawdźmy autokorelację pomiędzy każdytmi h-tymi residuami. Wyniki przedstawiony został na Wykresie 12. Wykres ten zdaje się potwierdzać tezę, że residua są nieskorelowane. W zerze funkcja autokorelacji wynosi 1, a w pozostałych punktach jest w okolicach zera, przy czym większość "słupków" nie wychodzi poza zaznaczony przedział ufności.



Wykres 12: Autokorelacja residuów

Sam wykres nie wystarczy do stwierdzenia czy residua są nieskorelowane. Przeprowadźmy test Ljung-Boxa, który służy do sprawdzenia czy dane są nieskorelowane. Wykorzystajmy funkcję `accor_ljungbox` z pakietu `statsmodels`. Ustalmy nasze hipotezy:

H_0 : residua są nieskorelowane,

H_1 : residuą nie są nieskorelowane.

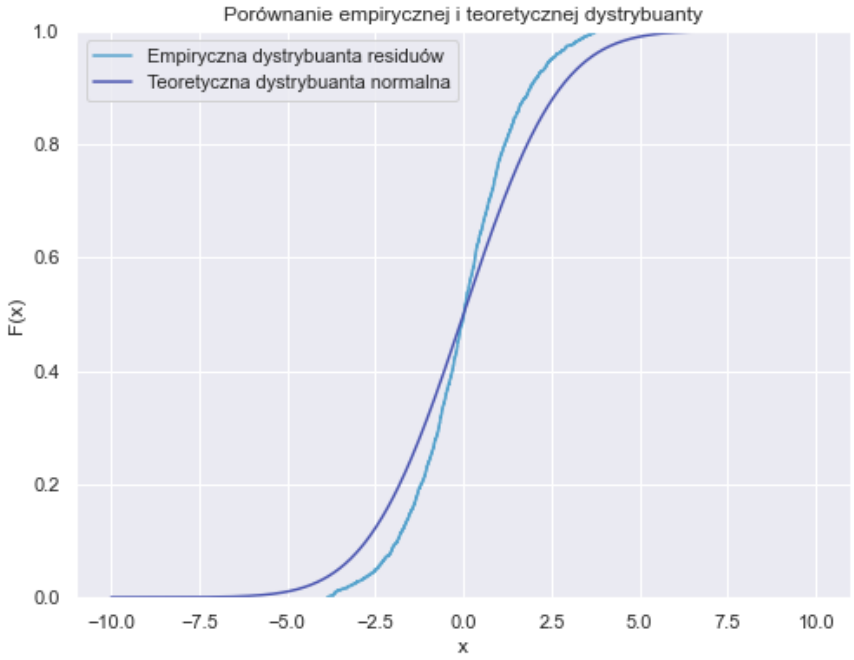
Po przeprowadzeniu testu, p-wartości osiągają wartości rzędów od 10^{-1} do 10^{-6}. Większość p-wartości jest o wiele mniejsza niż o poziom istotności α, przez co jesteśmy zmuszeni odrzucić hipotezę zerową i przyjąć, że residua nie są nieskorelowane.

4.4 Normalność residuów

Ostatnim krokiem naszej analizy będzie sprawdzenie, czy residua pochodzą z rozkładu normalnego.

4.4.1 Wizualizacja

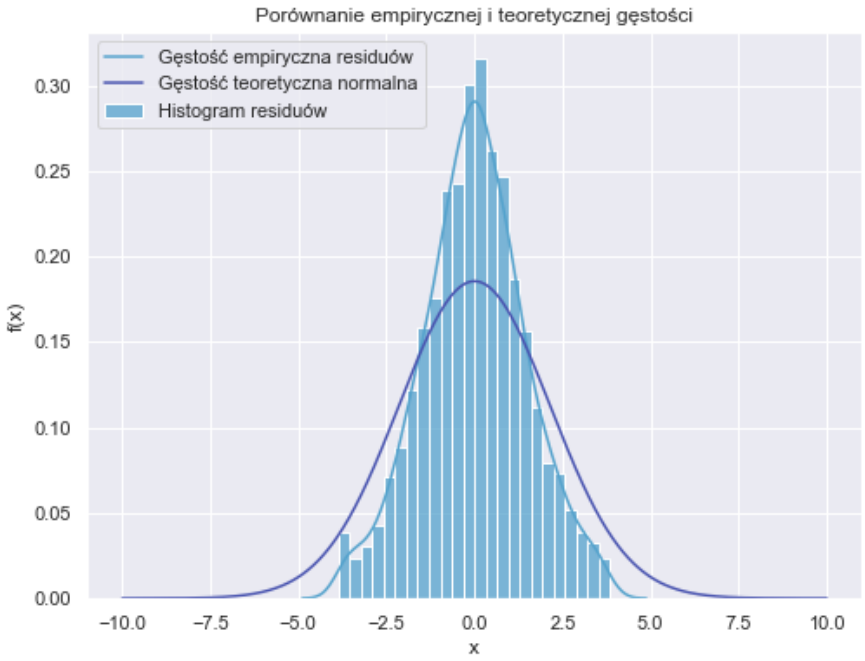
Zacznijmy od porównania dystrybuanty empirycznej residuów z teoretycznym rozkładem normalnym $\mathcal{N}(0,\, S^2)$. W tym celu skorzystamy z funkcji `ecdf` z biblioteki `seaborn` w Pythonie.



Wykres 13: Porównanie dystrybuanty empirycznej z teoretyczną

Jak widzimy na Wykresie 13 dystrybuanty nie pokrywają się.

Przejdźmy teraz do porównania gęstości empirycznej z teoretyczną. Do tej analizy użyliśmy funkcji `kdeplot` oraz `histplot` również z pakietu `seaborn`.



Wykres 14: Porównanie gęstości empirycznej z teoretyczną

Z Wykresu 14 możemy zobaczyć, że gęstości nie pokrywają się. W okolicach 0 występuje skok, który mówi nam o tym, że nasze błędy są często blisko 0.

4.4.2 Kurtoza, współczynnik skośności, reguła 3 sigm

Wzór na kurtozę i skośność można znaleźć w sekcji 2.2.3, natomiast w celu ich obliczenia użyliśmy funkcji wbudowanych `skew` oraz `kurtosis` z pakietu `scipy.stats`. Uzyskaliśmy, że kurtoza residuów jest równa -0.0315. Dla rozkładu normalnego powinna ona wynosić 0. Te liczby są zbliżone do siebie, czego można było się spodziewać po Wykresie 14.

W rozkładzie normalnym, przy $\mu = 0$ współczynnik skośności wynosi 0. Współczynnik skośności, w naszym przypadku wynosi -0.142. Jest to wartość bliska pożądanej.

Sprawdźmy teraz regułe trzech sigm. Wiemy, że dla zmiennej loswej $X \sim \mathcal{N}(\mu, \sigma^2)$ zachodzi reguła trzech sigm:

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.99 \tag{29}$$

Sprawdźmy czy nasze dane spełniają równanie 29.

$$ecdf(3S) - ecdf(3S) \approx 0.99 \tag{30}$$

ecdf to empiryczna funkcja dystrybuanty wyliczana za pomocą funkcji `stats.norm.cdf` z pakietu `scipy.stats`. Otrzymana wartość wynosi 0.959, co nie powinno nas dziwić bo jak widać na Wykresie 14 gęstość jest mocno skoncentrowana wokół zera.

4.4.3 Testy statystyczne

W celu ostatecznego sprawdzenia czy residua pochodzą z rozkładu normalnego, przeprowadziliśmy dwa testy: test Kołmogorowa-Smirnova oraz test Shapiro-Wilka. Ustalmy poziomi istotności $\alpha = 0.05$ hipotezę zerową i alternatywną:

$$H_0 : \text{residua są z rozkładu normalnego,}$$

H_1 : residuą nie są z rozkładu normalnego.

Do przeprowadzenia testów użyliśmy funkcji `kstest` oraz `shapiro` z biblioteki `scipy.stats`. Dla ks-testu p-wartość jest równa $1.3009 \cdot 10^{-12}$, a dla Shapiro-Wilka 0.0018. Obie te wartości są za małe. Jesteśmy zmuszeni odrzucić hipotezę zerową i przyjąć alternatywną.

4.5 Interpretacja wyników

Po przeanalizowaniu residuów wysnuliśmy wniosek, że nie pochodzą one z rozkładu normalnego. Pomimo tego, że ich wartość oczekiwana jest zbliżona do zera, wszystkie pozostałe założenia nie zostały spełnione. Ważne, żeby obserwacje wizualne potwierdzić obliczeniami czy też testami statystycznymi. Po Wykresie 12 moglibyśmy stwierdzić, że residua są niezależne, natomiast dopiero test statystyczny pokazał nam, że residua nie są nieskorelowane. Możemy stwierdzić, że wybrany model nie jest w pełni poprawny, mimo tego, że dane układają się w niemalże idealną prostą.

5 Podsumowanie i wnioski

W raporcie przeprowadzono analizę danych dotyczących cen akcji otwarcia i zamknięcia firmy The Walt Disney Company, która jest amerykańską korporacjaą środków masowego przekazu i rozrywki. Celem analizy było zbadanie zależności liniowej między dwoma zmiennymi - ceną otwarcia i ceną zamknięcia.

Analiza jednowymiarowa sugeruje, że można dobrze dopasować prostą regresji do analizowanych danych. Różnice między ceną otwarcia a zamknięcia są niewielkie, a wykresy obu zmiennych pokrywają się dobrze, co można zaobserwować na Wykresie 1. Zgodnie z oczekiwaniami, oczekiwana zależność zbliżona jest do funkcji $y = x$.

Pierwszym etapem analizy była porównawcza analiza jednowymiarowa, obejmująca gęstości, dystrybuanty i wykresy pudełkowe. Wykresy te pokazały, że dane są niemal identyczne, a różnice są minimalne, co potwierdzają również wartości statystyk.

Następnie wyznaczono wielkości statystyczne dla obu zmiennych. Parametry te różniły się jedynie kilkoma cyframi po przecinku, co wskazuje na podobieństwo danych. Warto zauważyć, że wykresy gęstości i dystrybuant oraz wykresy pudełkowe znacznie odbiegają od rozkładu normalnego.

Analiza zależności liniowej pomiędzy zmienną zależną i niezależną została zwizualizowana na Wykresie 5, prezentując prawie idealną zależność liniową. Współczynniki regresji liniowej, zarówno przed, jak i po usunięciu wartości odstających, potwierdziły oczekiwania.

Przed usunięciem wartości odstających otrzymano $\beta_0 = 0.3711$ i $\beta_1 = 0.9962$, a $R^2 = 0.996$. Po detekcji i usunięciu wartości odstających współczynniki te wyniosły odpowiednio $\beta_0 = 0.3422$ i $\beta_1 = 0.9964$, a $R^2 = 0.997$. Wzrost współczynnika R^2 po usunięciu wartości odstających świadczy o sukcesie w identyfikacji i eliminacji obserwacji odstających.

Zostały wyliczone przedziały ufności dla współczynników regresji. Współczynniki β_0 i β_1 znajdują się w przedziałach ufności. Warto zauważyć, że znajdowały się w nich nawet przed usunięciem wartości odstających.

Przeprowadziliśmy analizę residuów. Test Kołmogorowa-Smirnowa oraz Shapiro-Wilka pkazały nam, że residua nie pochodzą z rozkładu normalnego. Przy pomocy ttestu przyjęliśmy hipotezę, że wartość średnia residuów jest równa 0. Test Breusha-Pagana pomógł nam udowodnić, że wariancja nie jest stała, a za pomocą testu Ljung-Boxa, pokazaliśmy, że resiuda nie są nieskorelowane.

Resiuda spełniają tylko jedno założenie modelu regresji liniowej, jednakże bardzo dobrze dopasowana prosta regresji, wysoki wskaźnik R^2 nie pozwala nam przekreślić tego modelu.

Wyniki, które uzyskaliśmy, sugerują, że dokonaliśmy skutecznego doboru danych do modelu regresji liniowej. Przeprowadzona analiza jednowymiarowa, badanie zależności liniowych między zmiennymi oraz analiza reszt zostały przeprowadzone zgodnie z procedurą. Otrzymane rezultaty w większości przypadków harmonizują z naszymi przewidywaniami i intuicją.

Cała analiza przyczynia się do pogłębienia naszego zrozumienia dynamiki rynkowej firmy Disney, co z kolei może znacząco poprawić naszą zdolność do efektywnej oceny sytuacji na rynku kapitałowym.