SAD 2025L - Część 2 Projektu - Raport

Zespół nr 14 - Michał Piotrak (269336)

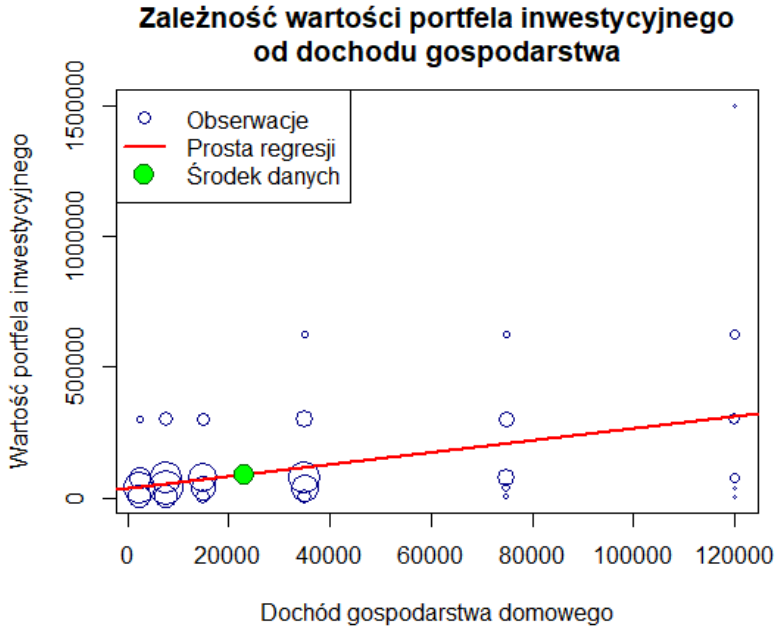
Krótka adnotacja a propos realizacji projektu
Projekt miał być realizowany w zespole dwuosobowym, wraz z Jakubem Zielińskim. Niestety, wraz z początkiem realizacji projektu kolega przestał odpowiadać na moje wiadomości. Ostatni raz kontaktowaliśmy się 28 maja 2025 r. Mieliśmy ustalony podział pracy i realizowałem projekt zgodnie z ustalonym planem, informując kolegę Jakuba o postępach, lecz on nie dawał żadnych sygnałów życia. Postanowiłem czekać do końcowego terminu oddania projektu, mając nadzieję, że kolega się odezwie, ale nie odezwał przez co cały projekt zrealizowałem sam. Niestety, do problemu nr 1 podszedłem już trochę wyrywkowo, ze względu na brak czasu i na to, że ten problem miał być realizowany przez kolegę Jakuba. Ze względu na zaistniały fakt, prosiłbym o mniej rygorystyczne ocenianie.

Problem 1

Aby zweryfikować przedstawioną w treści zadania zależność, że gospodarstwa domowe o większym dochodzie są bardziej skłonne do inwestowania większych kwot pieniędzy, postanowiłem wykorzystać model regresji liniowej. W celu jej użycia musiałem odpowiednio przetransformować dane dotyczące przedziałów określających zarówno miesięczny dochód gospodarstwa domowego oraz wartość portfela inwestycyjnego. Precyzując, postanowiłem te przedziały reprezentować poprzez ich środki, tj.:

- dla dochodów - (2500, 7500, 15000, 35000, 75000, 120000),
- dla wartości portfela - (5000, 40000, 75000, 300000, 625000, 1500000).

Liczbę osób podaną dla przecięcia tych dwóch rodzajów przedziałów, potraktowałem jako wartości wag dla otrzymanych obserwacji. W ten sposób otrzymaliśmy model regresji liniowej ważonej. Poniżej umieszczam wykres przedstawiający przebieg otrzymanej prostej regresji, z równoczesnym pokazaniem, że przechodzi ona przez środek ciężkości danych.



Odnośnie powyższego wykresu warto jeszcze dodać, że rozmiar kółka ma odzwierciedlać wagę danej obserwacji.

Otrzymany model regresji liniowej można przedstawić w postaci:
$Wartość\ portfela\ inwestycyjnego\ = 36878 + 2.305\ *\ Dochód$

Dla współczynnika *Dochód* otrzymano następujące wyniki, jeśli chodzi o jakość oszacowania oraz test istotności:

Nazwa parametru	Wartość parametru	Interpretacja wyniku
Błąd standardowy	0.822	Umiarkowana wartość, akceptowalna jakość estymacji
t value	2.804	Ta wartość oznacza, że dany współczynnik jest oddalony od 0 o 2.804 odchyleń standardowych - to gwarantuje dużą istotność tego współczynnika
p - value	0.00941	Tutaj dość podobnie jak w przypadku t value - mamy małą wartość, która wskazuje na duża istotność statystyczną współczynnika

Można jeszcze sprawdzić jakość modelu, jeśli chodzi o przewidywane wyniki i tutaj uzyskano takie wyniki:

- `Multiple R-squared:  0.2322`
- `Adjusted R-squared:  0.2027`

Wskazane tutaj wartości współczynnika determinacji R^2 są dość niskie, lecz może być to związane z tym, że na wartość portfela inwestycyjnego może wpływać jeszcze wiele innych czynników, nie tylko dochód gospodarstwa domowego.

Dla współczynnika *Dochód* zweryfikowałem jeszcze przedział ufności, który wyniósł *[0.616; 3.995]*. Dany przedział nie zawiera 0, co oznacza, że wpływ dochodu na wartość portfela jest istotny.

Podsumowując, uzyskany model regresji wykazał, że na wiekszą wartość portfela inwestycyjnego istotny wpływ ma miesięczny dochód, jaki generuje dane gospodarstwo domowe.

Problem 2

Do rozwiązania problemu opisanego w danym zadaniu zdecydowałem się wykorzystać dane dotyczące wartości akcji spółki strategicznej KGHM Polska Miedź S.A. w okresie 1 stycznia 2015 - 31 grudnia 2024 r. Wybór był podyktowany faktem, że spółka prowadzi politykę informowania inwestorów co kwartał o swoim wyniku finansowym, dlatego na wykresie wartości akcji tejże spółki można spodziewać się pewnych cyklicznych tendencji, które mogą być przydatne w ramach realizacji danego zadania. Dane pobrałem z portalu stooq.pl jako plik o formacie *.csv*, który został dołączony do rozwiązania zadania (plik o nazwie *kgh_m.csv*). Każdy wiersz w tym pliku zawiera informacje o wyniku otwarcia i zamknięcia danego miesiąca w okresie 1 stycznia 2015 - 31 grudnia 2024 r. w wykonaniu spółki.

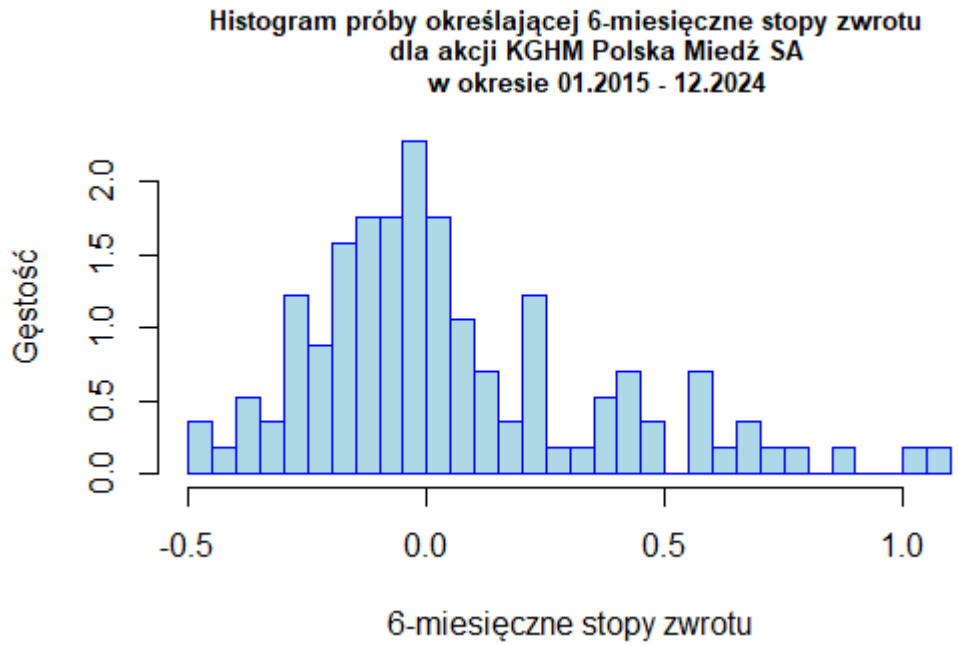
Całościowe rozwiązanie postawionego tutaj problemu zostało zawarte w pliku skryptowym o nazwie *Problem2.R*.

2.1 Strategia „od dziś na pół roku”

Zgodnie z poleceniami w treści tego podpunktu, wyznaczyłem szereg czasowy dotyczący wartości 6 - miesięcznych stóp zwrotu dla dowolnego momentu początku inwestycji w okresie styczeń 2015 - czerwiec 2024 (z tego względu że mamy dane do grudnia 2024, ostatnim możliwym do rozpatrzenia przeze mnie momentem początku inwestycji jest właśnie czerwiec 2024 r.).

Histogram

Poniżej został umieszczony histogram dla wyżej wymienionej próby:



Dobranie rodziny rozkładów

Analizując wygenerowany histogram, zwróciłem uwagę, że rozkład opisujący dane powinien być lekko skośny w lewo i posiadać grube ogony. Z tego powodu zdecydowałem się na próbę dopasowania rozkładu z naszej próby do rozkładu t - Studenta.

Wyestymowanie parametrów rozkładu

Do estymacji parametrów rozkładu t - Studenta wywołałem funkcję *fitdistr* z pakietu *MASS*. Dana funkcja swoje działanie opiera o wykorzystanie do estymacji metodę największego prawdopodobieństwa. Warto tutaj także zwrócić uwagę, że w moim rozwiązaniu dostarczyłem do funkcji *fitdistr* startową wartość stopni swobody. Dzięki temu, proces estymacji przestał zwracać wartości nieokreślone (*NaN*) i stał się stabilniejszy.

Ostatecznie estymacja parametrów przyniosła następujące rezultaty (zaokrąglenie do 5 miejsca po przecinku):

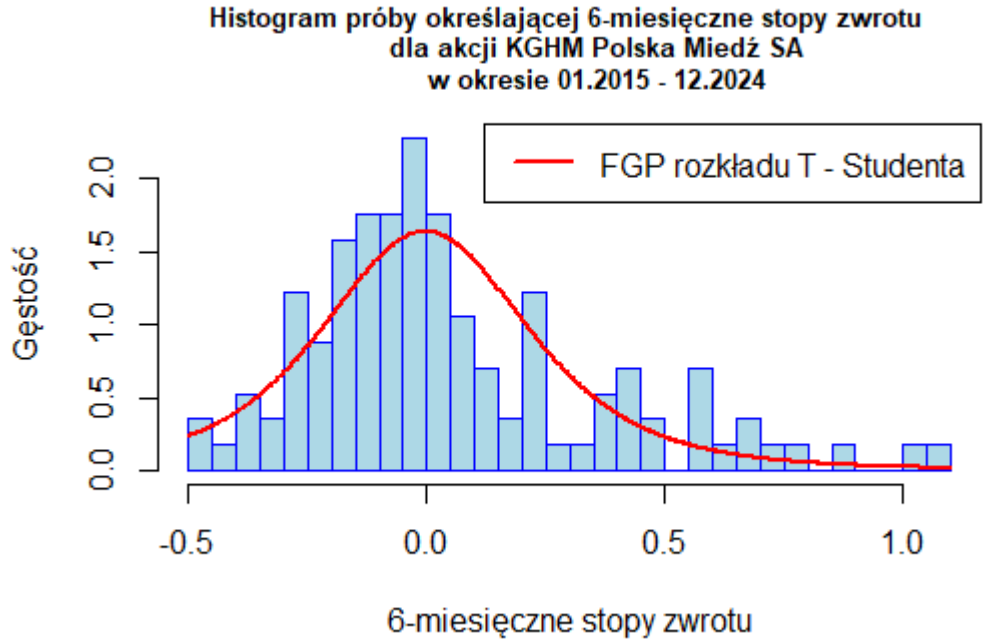
Nazwa parametru	**Wartość parametru**	**Odchylenie standardowe estymacji**
Położenie	-0,00356	0,03168
Skala	0,22753	0,03716
Stopnie swobody	3,62706	1,87514

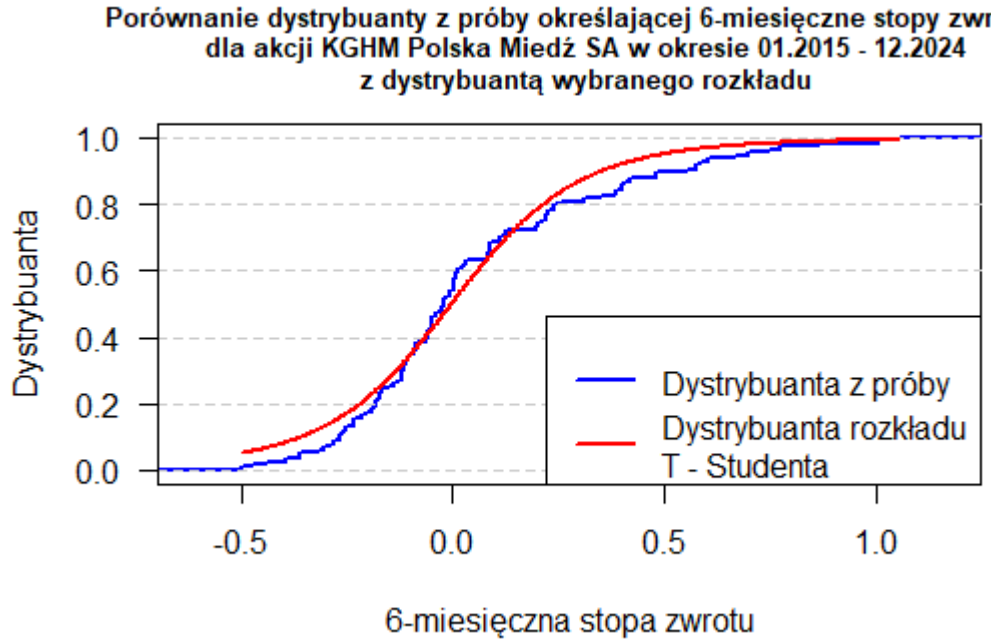
Z uzyskanych wartości parametrów możemy wywnioskować m.in.:

- wartość położenia bliska zeru sugeruje brak wyraźnej tendencji wzrostu / spadku,
- estymowana wartość stopni swobody wskazuje, że powinniśmy mieć więcej ekstremalnych przypadków niż w rozkładzie normalnym,
- duża wartość odchylenia standardowego przy estymacji parametru stopni swobody, więc duża niepewność estymacji.

Porównanie wygenerowanego rozkładu z wyestymowanymi parametrami z danymi

Moje badanie zgodności rozkładu próby do rozkładu t - Studenta z wyestymowanymi parametrami wyżej rozpocznę od części wizualnej. Poniżej zamieszczam porównanie histogramu próby do funkcji gęstości prawdopodobieństwa modelującego rozkładu oraz także porównanie dystrybuanty empirycznej z dystrybuantą teoretyczną rozkładu t - Studenta:

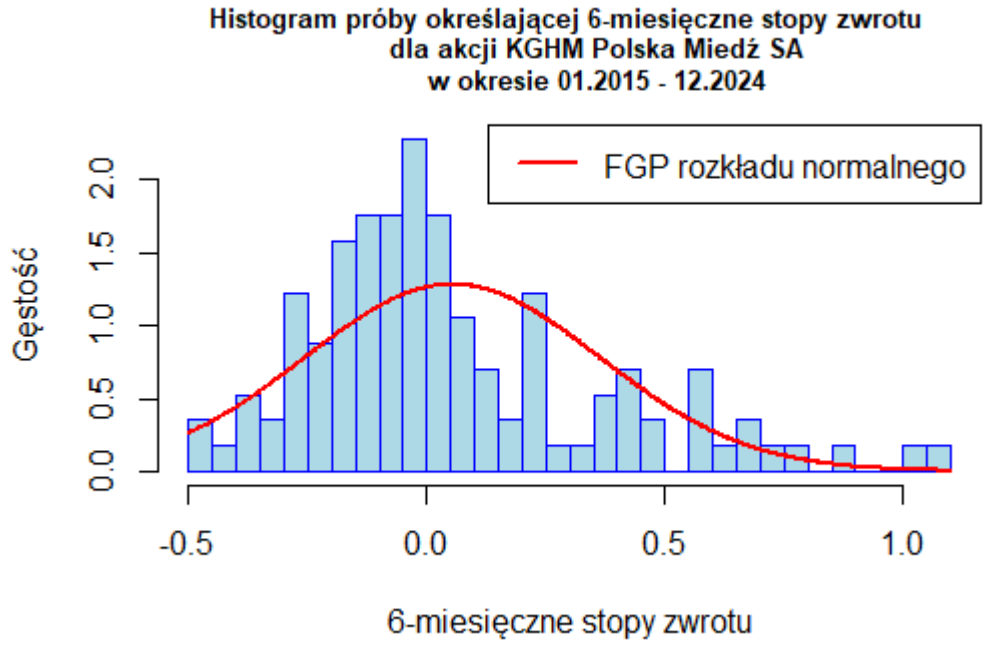


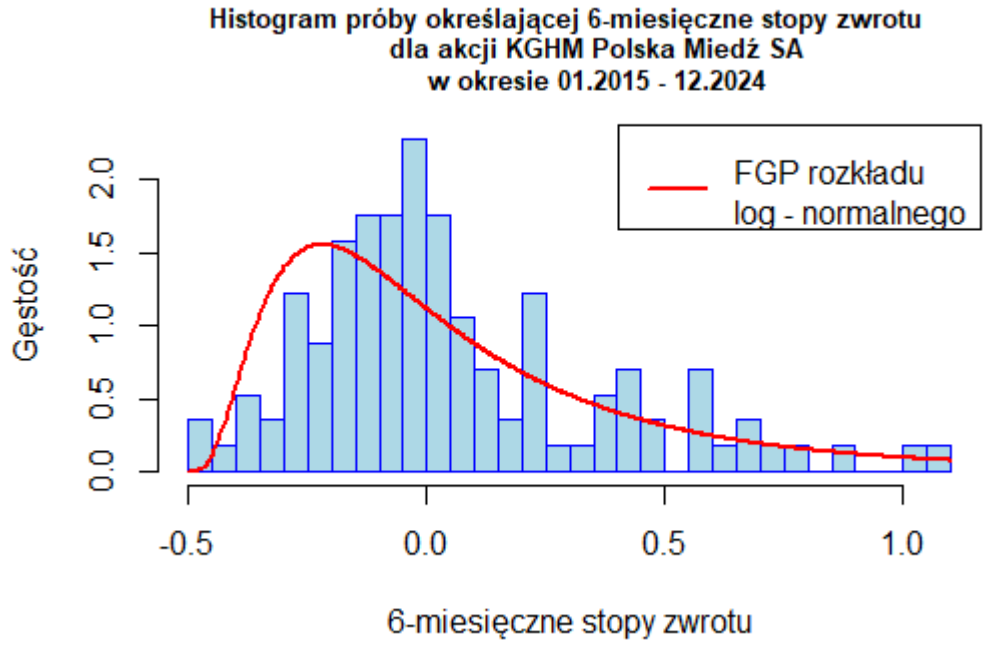


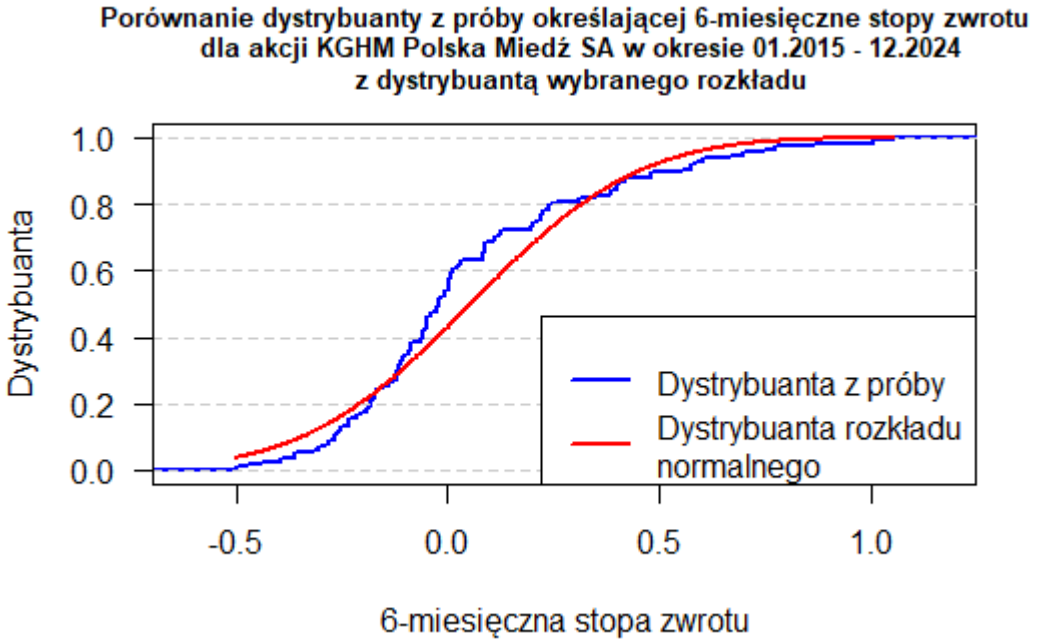
Przedstawione wyżej wykresy napawają optymizmem, jeśli chodzi o pozytywne rozstrzygnięcie sądów na temat zgodności danych do modelującego rozkładu. Na podstawie porównania histogramu do funkcji gęstości prawdopodobieństwa rozkładu t - Studenta możemy stwierdzić, że w tym przypadku mamy dobre dopasowanie do ogonów rozkładu oraz także sama dystrybuanta rozważanego rozkładu zachowuje należyte dopasowanie do dystrybuanty empirycznej.

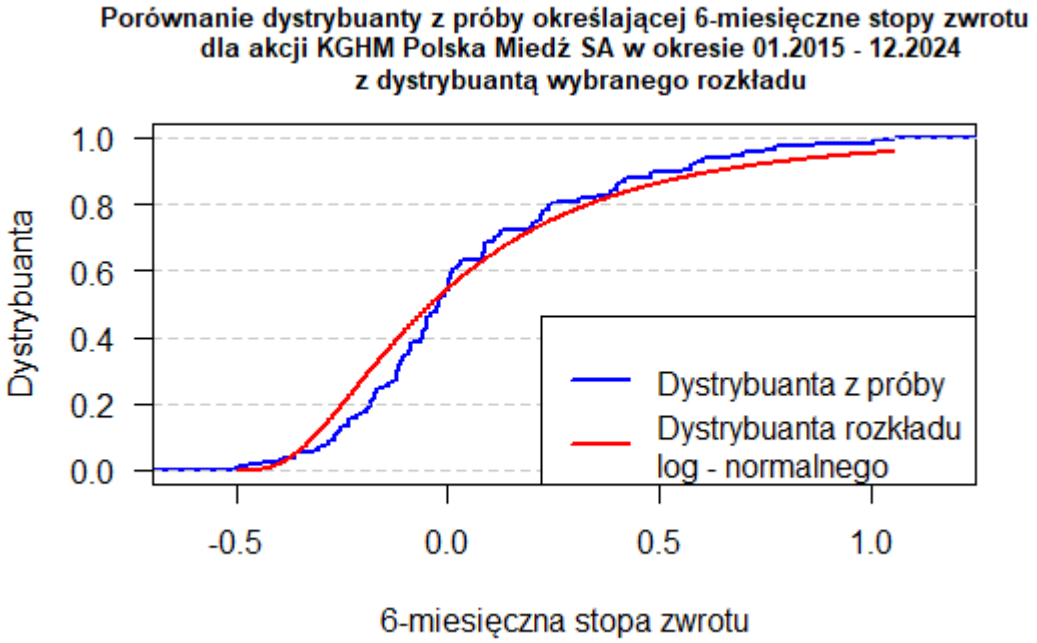
Dla pogłębienia analizy postanowiłem podobne rozważania przeprowadzić dla dwóch innych rozkładów - normalnego oraz logarytmiczno normalnego (dla którego musieliśmy dokonać odpowiedniego przesunięcia danych). Poniżej zamieszczam otrzymane wyniki, gdzie

możemy stwierdzić, że zaproponowane rozkłady nie gwarantują lepszego dopasowania do danych, niż rozkład t - Studenta:









Oczywiście, sprawdzenie wizualne nie jest wystarczające dla całego procesu dowodzenia zgodności rozkładu z danymi, dlatego został przeprowadzony odpowiedni test statystyczny, a dokładnie test Kołmogorowa - Smirnowa. Aby go wykonać musiałem odpowiednio przeskalować rozkład, uwzględniając jego położenie i skalę. Oto uzyskany rezultat (podane wartości w zaokrągleniu do 5 cyfry po przecinku):

Element wyniku testu	Opis elementu	Wartość
D	Określa największą różnicę między dystrybuantą	0,08878

	empiryczną a teoretyczną.	
p-value	Pozwala określić, czy hipoteza *H0* jest do odrzucenia, zakładająca, że nasze dane pochodzą z rozkładu t - Studenta.	0,33

Wartość *D* określa największą różnicę między dystrybuantą empiryczną a teoretyczną. Wartość ta wydaje się być dość mała, co sprzyja hipotezie *H0*. Jeszcze ważniejszym elementem tego testu jest *p-value*, której duża wartość tutaj oznacza brak podstaw do odrzucenia hipotezy *H0*. Dlatego, biorąc pod uwagę zarówno wynik testu Kołmogorowa - Smirnowa oraz analizę wizualną przedstawiona wyżej, z dużą pewnością możemy uznać, że nasza próba jest zgodna z rozkładem t - Studenta z wyestymowanymi parametrami.

2.2 Własna strategia

Poprzez wprowadzenie własnej strategii inwestycyjnej, ostatecznie ograniczam zbiór momentów startu inwestycji, bazując na z góry zdefiniowanej regule, która powinna zwiększać szansę na dodatni zysk. Odnośniej danej reguły, starałem się zastosować inne niż zaproponowane w treści zadania podejście, lecz ostatecznie, chcąc przedstawić jak najlepsze wyniki, postanowiłem jednak przyjąć następującą strategię - nie rozpoczynamy inwestycji. gdy w trzech ostatnich notowaniach cena spadała. Po wdrożeniu danego warunku, z pierwotnej próby liczącej 114 obserwacji dotyczących 6 - miesięcznych okresów inwestycji, pozostały nam 102 obserwacje jako nowa próba, którą należy przeanalizować.

Analogicznie do rozwiązania zawartego w podpunkcie 2.1, dokonałem próby dopasowania nowych danych do rozkładu t - Studenta. Poniżej tabelka z wyestymowanymi parametrami rozkładu.

Nazwa parametru	**Wartość parametru**	**Odchylenie standardowe estymacji**
Położenie	0,01781	0,03778
Skala	0,25781	0,04169
Stopnie swobody	6,0899	5,20167

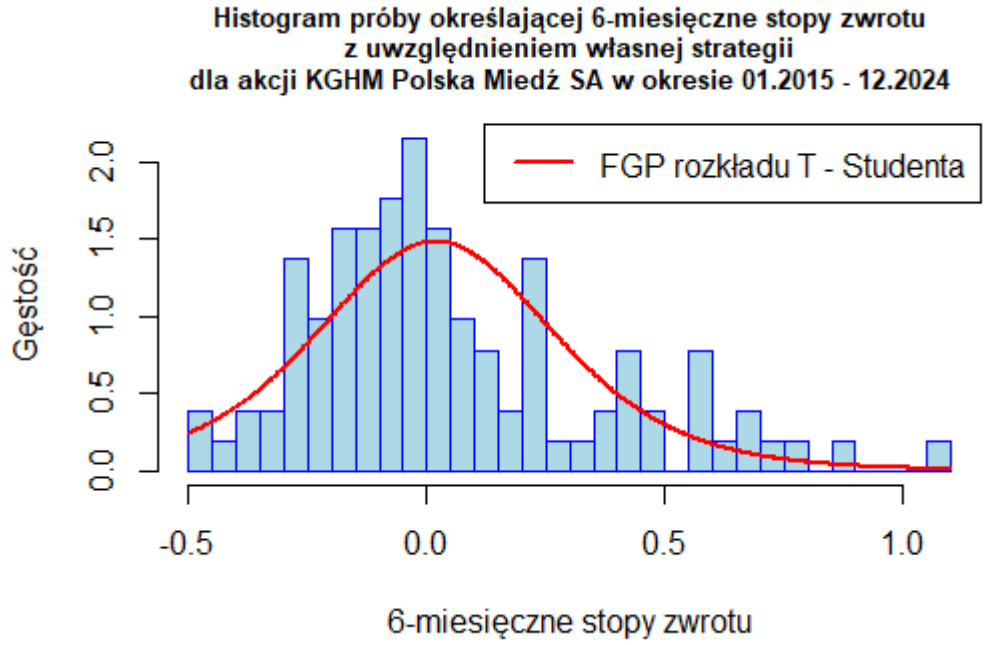
Dokonałem także porównania wartości średniej zysku oraz jego odchylenia standardowego dla tych dwóch prób:

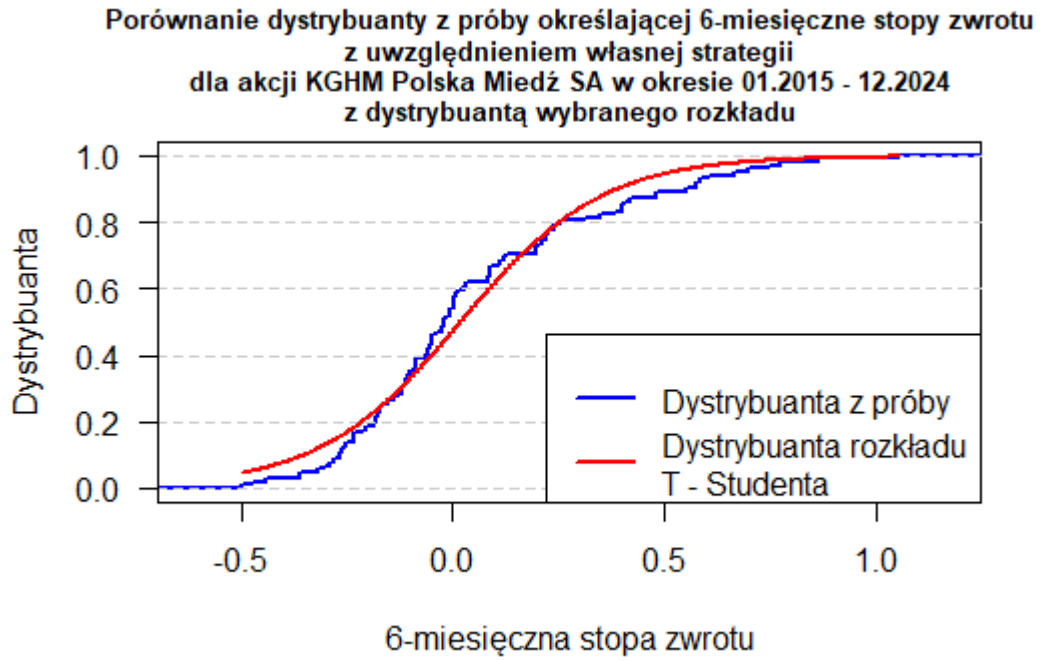
	Próba z strategią “od dziś na pół roku”	**Próba z własną strategią**
Średni zysk	0,05316	0,05384
Odchylenie standardowe	0,31303	0,3099

Z powyższych danych, wynikają następujące rzeczy:

- średni zysk delikatnie wzrósł na korzyść próby z zastosowaniem własnej strategii (o 0.07 p.p w zaokrągleniu),
- ryzyko inwestycji troszkę zmalało na korzyść próby z użyciem własnej strategii (o 0,3 p.p w zaokrągleniu).

Posiadając już wiedzę, że zastosowaną strategię można uznać za rozwiązanie co najmniej nie gorsze od podejścia standardowej 6 - miesięcznej inwestycji, przechodzimy do dalszego etapu analizy rozkładu z tejże nowej próby. Ponownie dokonałem wizualnej oceny porównawczej dopasowania danych do rozkładu t - Studenta i poniżej został zawarty wynik tej analizy:





Porównując z badaniem zgodności rozkładu w przypadku pierwotnej próby, uzyskujemy bardzo podobną zgodność ogólną dla nowej próbki, pomimo jej ograniczenia przez narzuconą strategię. Jedynie, na wykresie demonstrującym porównanie histogramu danych z funkcją gęstości prawdopodobieństwa rozkładu t - Studenta, można zauważyć delikatne pogorszenie w oszacowaniu lewego ogona. Lecz, dalej uzyskane dopasowanie można ocenić za dobre.

Analogicznie do podpunktu 2.1, wykonaliśmy test Kołmogorowa - Smirnowa. Poniżej przedstawiam jego wyniki:

Element wyniku testu	**Wartość**	**Interpretacja wyniku**
D	0,10771	Trochę większa różnica niż dla pierwszej próby, lecz dalej można ją uznać za wynik umiarkowany.
p-value	0,1875	Podobnie wynik gorszy niż w przypadku pierwszej próby, lecz dalej nie dostarcza powodów, żeby odrzucić hipotezę *H0*.

Podsumowując tę część badania rozkładu z nowej próby, zaproponowana strategia przyniosła delikatnie lepsze rezultaty, jeśli chodzi o wartość średniego zysku przy zachowaniu mniejszego ryzyka inwestycji. Potwierdziłem także jej zgodność z rozkładem t - Studenta.

Jak wcześniej wspomniałem, próbowałem zastosować inną strategię niż zaproponowaną w treści zadania. Pomyślałem o inwestowaniu wyłącznie w takich przypadkach, kiedy mamy po sobie 2 wzrosty z rzędu. Oto skondensowane wyniki wykorzystania takiego podejścia wraz z porównaniem do wyników otrzymanych z pierwotnej próby:

Miara	Próba z strategią “od dziś na pół roku”	Próba z strategią “2 wzrosty z rzędu”
Średni zysk	0,05316	0,03226
Odchylenie standardowe	0,31303	0,33213
Test KS - p-value	0,33	0,691

Jak widać po wynikach zawartych w tabelce, zastosowana tutaj strategia przyniosła gorszy średni zysk, przy delikatnie większym ryzyku. Jedynie, wynik testu Kołmogorowa - Smirnowa wskazuje, że ewentualnie dostaliśmy lepsze dopasowanie danych do rozkładu t - Studenta. Lecz, oczywiście strategia z zaniechaniem inwestowania przy 3 notowaniach spadkowych okazała się rozwiązaniem zdecydowanie skuteczniejszym, dlatego w pliku skryptowym zawierającym rozwiązanie, nie zawarłem kodu odtwarzającego dany eksperyment.