



Studium Magisterskie

Yauheni Semianiuk  
Nr albumu 82591

# **Analiza porównawcza metod grupowania na rynku akcji: perspektywa giełdy NASDAQ**

Warszawa 2024

# Spis treści

Wstęp .....	2
I. Teoretyczne podstawy analizy i grupowania akcji .....	4
1.1. Analiza akcji .....	4
1.2. Grupowanie w analizie finansowej.....	5
1.3. Opis giełdy NASDAQ.....	6
1.4. Przegląd literatury.....	7
II. Zbiór danych .....	9
2.1. Źródła i opis danych .....	9
2.2. Eksploracyjna analiza danych .....	11
III. Analiza empiryczna .....	16
3.1. Przebieg analizy.....	16
3.2. Wyniki porównania użytych metod.....	16
Zakończenie.....	27
Bibliografia.....	28
Spis rysunków .....	30
Spis tabeli .....	31

# Wstęp

Inwestycje finansowe są praktyką powszechnie stosowaną przez jednostki prywatne na całym świecie. Stanowią one kluczowe narzędzie w dążeniu do osiągnięcia różnorodnych celów, takich jak planowanie emerytalne, nabycie nieruchomości czy finansowanie edukacji dzieci. Osoby fizyczne decydują się na inwestycje w różne klasy aktywów, obejmujące akcje, obligacje, nieruchomości czy instrumenty finansowe o charakterze oszczędnościowym, a także nowoczesne formy, takie jak kryptowaluty czy crowdfunding. Te decyzje inwestycyjne pozwalają na wzrost kapitału, zabezpieczenie przed skutkami inflacji oraz generowanie pasywnego dochodu. W dzisiejszych czasach różnorodne instrumenty finansowe i platformy inwestycyjne stają się coraz bardziej dostępne dla szerokiego grona społeczeństwa, w efekcie sprzyjając elastycznemu oraz zindywidualizowanemu podejściu do inwestowania.

W świecie rozwijających się technologii i dostępności różnorodnych narzędzi inwestycyjnych, zrozumienie, jak efektywnie korzystać z tych informacji, staje się priorytetowym elementem dla osób indywidualnych. Dostęp do platform inwestycyjnych, analiz rynkowych online i doradców opartych o sztuczną inteligencję i zautomatyzowane algorytmy otwiera szerokie spektrum możliwości, ale jednocześnie niesie ze sobą potrzebę zwiększenia umiejętności w rozumieniu, analizie i podejmowaniu decyzji inwestycyjnych. W tym kontekście, jednostki indywidualne powinny rozwijać kompetencje w zakresie interpretacji danych finansowych, analizy trendów rynkowych oraz identyfikacji kluczowych wskaźników ekonomicznych. Umiejętność dokładnego zrozumienia informacji dostarczanych przez różnorodne źródła, takie jak raporty finansowe, analizy branżowe czy prognozy ekonomiczne, jest niezbędna do skutecznego oceniania perspektyw inwestycyjnych.

W przypadku rynków finansowych, akcje notowane na renomowanych giełdach zajmują wyjątkowe miejsce w świadomości inwestorów prywatnych. NASDAQ, będąc jedną z najbardziej aktywnych i dynamicznych giełd na świecie, pełni nadzwyczaj istotną rolę w globalnym krajobrazie finansowym, przyciągając uwagę inwestorów z różnych zakątków świata. To miejsce, gdzie dokonuje się transakcje na akcjach nie tylko dużych, ale również innowacyjnych firm, które kształtują przyszłość różnych sektorów gospodarki. Inwestorzy, poszukując udziału w rynku, gdzie dominuje wysoka płynność i intensywna aktywność handlowa, często kierują swoją uwagę właśnie na NASDAQ. Zróżnicowanie notowanych na NASDAQ spółek akcyjnych umożliwia inwestorom konstruowanie zbilansowanych portfeli, uwzględniających potencjał zarówno ugruntowanych przedsiębiorstw, jak i perspektywicznych startupów.

Niniejsza praca będzie skupiała się na analizie poszczególnych akcji notowanych na NASDAQ oraz zastosowaniu analizy skupień w celu utworzenia wyraźnych zgrupowań obiektów, składających się ze spółek, które wykazują podobne wzorce zachowań na rynku. Przeanalizowane zostaną zarówno fundamentalne aspekty, jak i techniczne wskaźniki rynkowe, obejmujące dynamikę cenową oraz trendy handlowe. Taka wielowymiarowa analiza umożliwi pełniejsze zrozumienie charakterystyki każdej z badanych spółek oraz poprawne zgrupowanie, odzwierciedlające kompleksowe relacje zachodzące pomiędzy spółkami notowanymi na NASDAQ. Z kolei głównym celem niniejszej pracy jest porównanie różnych metod grupowania w celu znalezienia odpowiedniej, biorąc pod uwagę wybrane czynniki, metody. Pozwoli to zidentyfikować najbardziej efektywną i precyzyjną metodą klasyfikacyjną, które pozwoli na zrozumienie powtarzających się wzorców, istotnych dla podejmowania decyzji inwestycyjnych.

W rozdziale I poświęcono czas wyjaśnieniu fundamentalnych podstaw teoretycznych analizy i grupowania cen akcji, bardziej szczególnego przedstawienia giełdy NASDAQ oraz przedyskutowano istniejącą literaturę badawczą. Rozdział II skupia się na opisie analizowanego zbioru danych oraz analizy eksploracyjnej. Z kolei Rozdział III pokazuje bezpośrednio analizę empiryczną, przedstawiając wyniki porównania wybranych metod. Praca kończy się podsumowaniem uzyskanych rezultatów, omówieniem potencjalnych zastosowań oraz kierunek przyszłych badań.

# I. Teoretyczne podstawy analizy i grupowania akcji

## 1.1. Analiza akcji

Analiza akcji stanowi kluczowy element w dziedzinie inwestycji, mający na celu zrozumienie i ocenę papierów wartościowych emitowanych przez spółki publiczne. Formalnie, analiza akcji to proces, w ramach którego analizowane są finansowe, ekonomiczne i rynkowe aspekty spółek notowanych na giełdzie. Celem tego procesu jest zrozumienie bieżącej sytuacji i przyszłego potencjału inwestycyjnego danej firmy, co umożliwia podejmowanie świadomych decyzji inwestycyjnych. Praktycznie, analiza akcji obejmuje zastosowanie różnych metod badawczych w celu gromadzenia, przetwarzania i interpretacji informacji dotyczących danej spółki, jej otoczenia rynkowego oraz ogólnego stanu gospodarki.

Do najbardziej powszechnych metod analizy akcji należą:

- Analiza fundamentalna: skupia się na ocenie fundamentalnych wskaźników finansowych i ekonomicznych spółki. Obejmuje analizę sprawozdań finansowych, bilansów, rachunków zysków i strat, a także ocenę potencjału wzrostu, skuteczności zarządzania i otoczenia makroekonomicznego.
- Analiza techniczna: koncentruje się na badaniu historycznych danych cenowych oraz wolumenu obrotu, aby identyfikować wzorce i trendy rynkowe. Wykorzystuje narzędzia graficzne i matematyczne do prognozowania przyszłych ruchów cenowych.

Analiza akcji opiera się na pewnych naukowych koncepcjach, z których najważniejsze to:

- Hipoteza rynku efektywnego: według tej teorii, ceny akcji odzwierciedlają wszystkie dostępne informacje, co utrudnia osiągnięcie stałych nadzwyczajnych zysków poprzez jedynie analizę rynku. Hipoteza ta podkreśla, że inwestorzy nie są w stanie systematycznie przewyższać rynku, bazując wyłącznie na analizie informacji publicznie dostępnych.
- Model rynku behawioralnego: model ten zakłada, że inwestorzy nie zawsze podejmują racjonalne decyzje oparte na informacjach, lecz są również podatni na emocje i uprzedzenia. Analiza akcji w oparciu o ten model uwzględnia psychologiczne aspekty inwestycji, takie jak efekt stadny czy tendencje do podejmowania decyzji na podstawie bieżących trendów rynkowych.
- Teoria portfela efektywnego: według tej teorii, inwestorzy starają się zoptymalizować swoje portfele, uwzględniając równocześnie zarówno oczekiwane zyski, jak i ryzyko. Analiza akcji

w świetle tej teorii koncentruje się na konstrukcji portfeli, które oferują maksymalny poziom zwrotów przy określonym poziomie ryzyka.

- Teoria opłacalności dla inwestora: teoria ta zakłada, że inwestorzy są skłonni ponosić dodatkowe koszty i trudy analizy akcji tylko wtedy, gdy potencjalne korzyści przewyższają koszty transakcyjne i wysiłek związany z badaniami. Analiza akcji w oparciu o tę teorię bada, jakie czynniki wpływają na decyzje inwestycyjne, uwzględniając koszty i korzyści analizy.

Z kolei znaczenie analizy akcji w dziedzinie inwestowania polega przede wszystkim na:

- Ocenie wartości firmy: analiza akcji pełni kluczową rolę w procesie oceny rzeczywistej wartości firmy, gdyż umożliwia bardziej świadome podejmowanie decyzji inwestycyjnych, opartych na solidnych fundamentach ekonomicznych.
- Zarządzanie ryzykiem: analiza akcji służy również jako narzędzie zarządzania ryzykiem poprzez identyfikację potencjalnych czynników wpływających na wartość akcji. W rezultacie inwestorzy są w stanie lepiej dostosować swoje strategie inwestycyjne do panujących warunków rynkowych, minimalizując ryzyko i zwiększając szanse na osiągnięcie pożądaných rezultatów.

W ten sposób analiza akcji to kompleksowy proces, który obejmuje różne metody badawcze i koncepcje naukowe. Wnikliwe zrozumienie tego procesu jest kluczowe dla skutecznego zarządzania portfelem inwestycyjnym.

## **1.2. Grupowanie w analizie finansowej**

Grupowanie, znane również jako klasyfikacja, w kontekście analizy finansowej odnosi się do procesu dzielenia zbioru danych na różne kategorie lub grupy na podstawie pewnych wspólnych cech, lub kryteriów. Celem tego procesu jest zidentyfikowanie wzorców, relacji lub trendów w danych, co może pomóc w lepszym zrozumieniu struktury danych i ułatwieniu podejmowania decyzji. Jedną z najważniejszych technik w tym obszarze jest analiza skupień, która polega na podziale danych na grupy na podstawie podobieństw między nimi. W kontekście analizy finansowej analiza skupień może wspomagać identyfikację podobieństw między różnymi spółkami na podstawie ich charakterystyk finansowych. Na przykład, spółki o zbliżonych wskaźnikach rentowności czy zadłużenia mogą być przypisane do jednego klastra.

Grupowanie w analizie akcji przynosi szereg korzyści, które znacznie ułatwiają zrozumienie i interpretację danych finansowych oraz wspierają podejmowanie decyzji inwestycyjnych. Oto kilka głównych korzyści wynikających z zastosowania grupowania w analizie akcji:

- Identyfikacja podobieństw i różnic: grupowanie pozwala na wyodrębnienie podobieństw między różnymi akcjami na podstawie ich charakterystyk finansowych. Identyczność pewnych cech może wskazywać na wspólne trendy czy czynniki wpływające na wartość tych akcji. Z drugiej strony, grupowanie pozwala również na zauważenie różnic, co jest istotne dla zrozumienia unikalnych cech każdej spółki.
- Ułatwienie analizy porównawczej: grupy obiektów, utworzone za pomocą grupowania, ułatwiają porównania między różnymi spółkami. To przydatne narzędzie dla tych, którzy poszukują spółek o podobnych profilach inwestycyjnych.
- Optymalizacja dywersyfikacji portfela: grupowanie pozwala inwestorom na lepsze zrozumienie struktury swojego portfela inwestycyjnego. Identyfikacja różnych klastrów może sugerować, w jakie sektory czy branże inwestorzy są obecnie zaangażowani. Pozwala to na optymalizację dywersyfikacji portfela, minimalizując ryzyko związane z koncentracją w jednym obszarze.
- Wsparcie w procesie decyzyjnym: grupowanie dostarcza jasnych i zorganizowanych informacji, co ułatwia podejmowanie decyzji inwestycyjnych. Inwestorzy mogą skoncentrować się na analizie charakterystyk danego klastra, co ułatwia wybór spółek zgodnych z określonym profilem inwestycyjnym.
- Analiza trendów rynkowych: grupowanie pozwala na zidentyfikowanie trendów rynkowych, takich jak dominujące sektory czy branże. Inwestorzy mogą lepiej zrozumieć, które akcje są obecnie atrakcyjne w danym kontekście rynkowym, co może być przydatne dla dostosowania strategii inwestycyjnej.

### 1.3. Opis giełdy NASDAQ

Giełda NASDAQ<sup>1</sup> (z ang. *National Association of Securities Dealers Automated Quotations*) to jedna z najważniejszych giełd papierów wartościowych na świecie, odgrywająca kluczową rolę w globalnym krajobrazie finansowym. Jej historia sięga roku 1971, kiedy została założona jako pierwsza giełda, która wprowadziła elektroniczny system notowań. Innowacyjny charakter NASDAQ od samego początku wpłynął na sposób, w jaki odbywają się transakcje na rynkach finansowych. NASDAQ wyróżnia się przede wszystkim jako miejsce notowań dla przedsiębiorstw z sektora technologicznego. Giełda ta również zyskała renomę jako platforma, na której notowane są akcje nie tylko dużych, ale także stosunkowo niewielkich i dynamicznych przedsiębiorstw, które kształtują

---

<sup>1</sup> <https://www.nasdaq.com/>, dostęp 20.01.2024.

przyszłość różnych sektorów gospodarki. Specyfika NASDAQ polega na tym, że umożliwia notowanie spółek, które często charakteryzują się wysokim tempem rozwoju, a także poszukują kapitału na ekspansję na rynku.

W takim dynamicznym i często ryzykownym kontekście, szczególnie istotne jest jak najlepsze zrozumienie relacji pomiędzy różnymi przedsiębiorstwami notowanymi na NASDAQ. Ze względu na zróżnicowanie sektorów gospodarki, w których działają te przedsiębiorstwa, a także odmienności w stopniu innowacyjności czy dynamice ich działań, skrupulatne wyodrębnienie grup spółek staje się kluczowym elementem analizy. Firmy notowane na NASDAQ są bardzo zróżnicowane zarówno pod względem modelu biznesowego, jak i profilu ryzyka. Dlatego skuteczna identyfikacja i zgrupowanie podobnych podmiotów umożliwia inwestorom bardziej precyzyjne podejście do zarządzania portfelem inwestycyjnym, licząc się z specyfiką każdej grupy spółek.

## 1.4. Przegląd literatury

Istnieje sporo badań skupiających się na stosowaniu różnych technik grupowania. Bini i Tessy porównywali różne techniki grupowania<sup>2</sup> i wykazali, że najwyższą wydajność osiąga się w przypadku algorytmu K-średnich oraz algorytmu Expectation-Maximization. Z kolei Dattatray, Gandhmal i Kumar przeanalizowali 50 artykułów naukowych,<sup>3</sup> podkreślając, że powszechnie stosowaną techniką grupowania akcji są modele sztucznych sieci neuronowych. Zheng i Chaoshin zbadali zastosowanie algorytmu grupowania K-średnich w rekomendacjach akcji,<sup>4</sup> wykazując, że 100 akcji można skutecznie podzielić na te o wysokich wynikach i te o słabych wynikach poprzez grupowanie. Al-Augby, Majewski, Majewska i Nermend porównali grupowanie K-średnich i rozmytych K-średnich,<sup>5</sup> sugerując użycie rozmytych K-średnich jako statycznie efektywniejszej metody grupowania. Jednakże wyniki analizy Nanda, Mahanty i Tiwari pokazują, że analiza skupień K-średnich pozwala zbudować bardziej zwarte klastry w porównaniu do metody rozmytych średnich dla klasyfikacji akcji.<sup>6</sup>

---

<sup>2</sup> Bini B.S., Tessy M., *Clustering and Regression Techniques for Stock Prediction*, Procedia Technology Wyd. 24, 2016, s. 1248-1255.

<sup>3</sup> Dattatray D., Gandhmal P., Kumar K., *Systematic analysis and review of stock market prediction techniques*, Computer Science Review Wyd. 34, 2019, s. 1000-1019.

<sup>4</sup> Zheng F., Chaoshin C., *Research on Prediction and Recommendation of Financial Stocks Based on K-means Clustering Algorithm Optimization*, Journal of Computational Methods in Sciences and Engineering, Wyd. 21 Nr. 5, 2021, s. 1081 – 1089.

<sup>5</sup> Al-Augby S., Majewski S., Majewska A., Nermend K., *A Comparison Of K-Means and Fuzzy C-Means Clustering Methods for A Sample of Gulf Cooperation Council Stock Markets*, Folia Oeconomica Stetinensia, Wyd. 14 Nr. 2, 2014, s. 19-36.

<sup>6</sup> Nanda S. R., Mahanty B., Tiwari M. K., *Clustering Indian stock market data for portfolio management*, Expert Systems with Applications Wyd. 37 Nr. 12, 2010, s. 8793-8798.



Jednakże warto zaznaczyć, że do chwili obecnej brakuje badań, które porównywałyby algorytm k-średnich z metodą Wisharta, metodą podziału dendrogramu, metodami deklomerycyjnymi lub metodą wrocławską. Co więcej, w większości dotychczasowych zastosowań algorytm k-średnich pełnił jedynie rolę pierwszego z etapów analizy, podczas gdy w niniejszym badaniu skoncentrowano się na bezpośrednim porównaniu tych metod i identyfikacji optymalnej techniki.

## II. Zbiór danych

### 2.1. Źródła i opis danych

Zestaw danych analizowany w niniejszym badaniu został pozyskany za pośrednictwem serwisu informacyjnego Yahoo Finance<sup>7</sup> oraz przez firmę tradingową Optiver.<sup>8</sup> Obejmuje on różne charakterystyki 196 akcji notowanych na giełdzie NASDAQ w roku 2023. Analizowany zbiór zawiera dwie kategorie zmiennych, reprezentujących czynniki z podejścia analizy fundamentalnej oraz z podejścia analizy technicznej. Zmienne z pierwszej kategorii zawarte w zbiorze obejmują:

- *at* – wydajność aktywów (ang. *asset turnover*), która mierzy, jak efektywnie firma generuje przychody z posiadanych aktywów w danym okresie. Jest obliczana jako stosunek przychodów ze sprzedaży do średnich aktywów firmy. Wyższa wartość sugeruje wydajniejsze wykorzystanie aktywów w generowaniu przychodów.
- *cdpr* – stosunek wypłaconych dywidend gotówkowych (ang. *cash dividends paid ratio*). Jest to wskaźnik określający, jaką część zysków spółki stanowią wypłacone dywidendy w gotówce. Jest obliczany jako stosunek wypłaconych dywidend do zysku netto.
- *cftc* – przepływ gotówkowy do wydatków na kapitał trwały (ang. *cash flow to capex*). Jest to miara oceniająca zdolność firmy do generowania wystarczającego przepływu gotówkowego, aby pokryć wydatki na inwestycje w kapitał trwały. Wartość powyżej jedności sugeruje zdolność firmy do samofinansowania inwestycji.
- *current\_ratio* – stosunek bieżący, który mierzy zdolność firmy do pokrycia krótkoterminowych zobowiązań za pomocą aktywów krótkoterminowych. Wyższa wartość wskazuje na lepszą płynność finansową.
- *dte* – stosunek zadłużenia do kapitału własnego (ang. *debt to equity ratio*), który określa, jak dużo kapitału firmy pochodzi ze źródeł zadłużenia w porównaniu do kapitału własnego. Wyższa wartość może wskazywać na większe ryzyko związane z zadłużeniem.
- *eps* – zysk na akcję (ang. *earnings per share*), czyli miara przedstawiająca część zysku przypisaną jednej akcji. Wskaźnik ten jest często używany do oceny rentowności akcji.
- *fcftni* – dostępny przepływ gotówkowy do dochodu netto (ang. *free cash flow to net income*). Jest to wskaźnik oceniający, jaką część zysku netto generuje firma w formie wolnych środków pieniężnych.

---

<sup>7</sup> <https://finance.yahoo.com/>, dostęp 20.01.2024.

<sup>8</sup> <https://www.kaggle.com/competitions/optiver-trading-at-the-close/overview>, dostęp 20.01.2024.

- gpm – marża zysku brutto (ang. *gross profit margin*), czyli procentowy udział zysku brutto w przychodach ze sprzedaży. Wyższa wartość sugeruje lepszą rentowność operacyjną.
- ocfts – przepływ gotówkowy operacyjny do sprzedaży (ang. *operating cash flow to sales*), mówiący o tym, jak dużo gotówki generuje się z działań operacyjnych w porównaniu do przychodów ze sprzedaży.
- oip – procentowy udział zysku operacyjnego (ang. *operating income percentage*). Jest to miara przedstawiająca procentowy udział zysku operacyjnego w przychodach ze sprzedaży. Wyższa wartość sugeruje lepszą rentowność operacyjną.
- q\_ratio – stosunek szybki (ang. *quick ratio*). Wskaźnik podobny do stosunku bieżącego, ale uwzględniający jedynie najbardziej płynne aktywa krótkoterminowe. Mierzy zdolność firmy do spłaty bieżących zobowiązań z najbardziej płynnych aktywów.
- rdip – procentowy udział wydatków na badania i rozwój (ang. *research and development investment percentage*), przedstawiająca procentowy udział wydatków firmy na badania i rozwój w przychodach ze sprzedaży. Wyższa wartość może wskazywać na większe zaangażowanie w innowacje.
- roa – zwrot z aktywów (ang. *return on assets*). Wskaźnik oceniający zdolność firmy do generowania zysków z posiadanych aktywów. Obliczany jest jako stosunek zysku netto do średnich aktywów. Wyższa wartość sugeruje efektywne wykorzystanie aktywów.
- roe – zwrot z kapitału własnego (ang. *return on equity*). Wskaźnik mierzący efektywność wykorzystania kapitału własnego do generowania zysku netto. Wyższa wartość może wskazywać na lepszą rentowność kapitału własnego.

Zmienne te reprezentują wskaźniki finansowe uzyskane przez cały poprzedni rok fiskalny. Z kolei do drugiej kategorii zmiennych, reprezentującej czynniki analizy technicznej należą:

- stmpp – procentowe odchylenie standardowe do ceny średniej (ang. *standard deviation to mean price percentage*). Wyższa wartość może wskazywać na większą zmienność cenową.
- lp\_ratio – stosunek najniższej ceny do średniej ceny dostosowanej (ang. *lower price ratio*). Wyższa wartość sugeruje, że cena akcji była niska w porównaniu do średniej ceny.
- up\_ratio – stosunek najwyższej ceny do średniej ceny dostosowanej (ang. *upper price ratio*). Wyższa wartość sugeruje, że cena akcji była stosunkowo wysoka w porównaniu do średniej ceny.
- dividends – wartość dywidend wypłacona w ciągu roku.

W przypadku opisanych powyżej zmiennych, używane były wartości średnie w ujęciu rocznym. Przed obliczeniem średniej, każda wartość reprezentowała konkretny dzień. Ostatnia część użytych zmiennych diagnostycznych dotyczy zmiennych wśród dziennych:

- `imb_size` – wielkość nierównowagi (*imbalance size*), czyli ilość niezrównoważonych ofert w stosunku do bieżącej ceny referencyjnej.
- `imb_flag` – wskaźnik odzwierciedlający kierunek nierównowagi aukcyjnej.
- `ref_price` – cena referencyjna, czyli cena, przy której ilość sparowanych akcji jest maksymalizowana, nierównowaga jest minimalizowana, oraz odległość od środka między najwyższą a najniższą ceną jest minimalizowana. Może być także traktowana jako cena bliska, ograniczona między najlepszym zleceniem kupna a sprzedaży.
- `matched_size` – sparowana wielkość. Ilość, która może być sparowana w bieżącej cenie referencyjnej.
- `far_price` – cena daleka, która zwiększa liczbę sparowanych akcji opartych tylko na zainteresowaniu aukcyjnym.
- `near_price` – cena bliska, która zwiększa liczbę sparowanych akcji opartych na aukcjach i ciągłych zleceniach rynkowych.
- `bid/ask_price` – cena kupna/sprzedaży, czyli cena najbardziej konkurencyjnego poziomu kupna/sprzedaży w księdze zleceń poza aukcją.
- `bid/ask_size` – wielkość kupna/sprzedaży, czyli kwota nominalna na najbardziej konkurencyjnym poziomie kupna/sprzedaży w księdze zleceń poza aukcją.
- `wap` – ważona średnia cena w księdze zleceń poza aukcją.
- `target` – przewidywane przesunięcie w przyszłości ceny średniej ważonej akcji w ciągu 60 sekund, pomniejszone o przesunięcie w przyszłości ceny średniej ważonej syntetycznego indeksu.

W przypadku opisanych powyżej zmiennych, używane były wartości średnie. Przed agregacją, każda obserwacja reprezentowała średnią wartość za każde 10 sekund z pięciu ostatnich minut otwartych sesji giełdowych w każdym dniu handlowym.

## 2.2. Eksploracyjna analiza danych

Proces eksploracyjnej analizy zgromadzonego zbioru danych rozpoczął się od oceny wartości informacyjnej poszczególnych zmiennych. W tym celu stworzono tabelę zawierającą współczynniki zmienności dla każdej zmiennej osobno.

Tabela 1. Współczynniki zmienności zmiennych diagnostycznych

Stats	ask_price	ask_size	asset_turn	bid_price	bid_size	cash_flow	cash_flow_x	current_ratio	debt_turn	dividend	eps	far_price
Mean	1.000216	72402.42	.0828578	.9997458	69637.37	-.1607629	9.312822	2.75682	-21.15193	1.343915	1.912053	1.001768
SD	.0002939	68153.42	.0628637	.0002246	67610.68	.356061	258.788	5.53516	385.9771	2.733959	7.442496	.0147306
CV	.0002939	.9413142	.7586934	.0002246	.9708966	-2.214821	27.78835	2.007806	-18.24784	2.034325	3.89241	.0147046

Stats	free_cash	gross_turn	imb_flag	imb_size	lower_ratio	match_turn	near_price	operat_turn	operat_turn	quick_ratio	rdi_perc	ref_price
Mean	.2300355	.4788709	-.0168669	2633346	.7281753	5.57e+07	.9996525	.6587828	.4537358	2.022613	.2395402	.9999822
SD	9.574806	.336098	.0932345	4438908	.1478408	1.11e+08	.0008246	10.32551	2.563752	4.102645	.2375936	.0001862
CV	41.62317	.7018551	-5.527673	1.685653	.2030291	1.998943	.0008249	15.67362	5.650319	2.028389	.9918736	.0001862

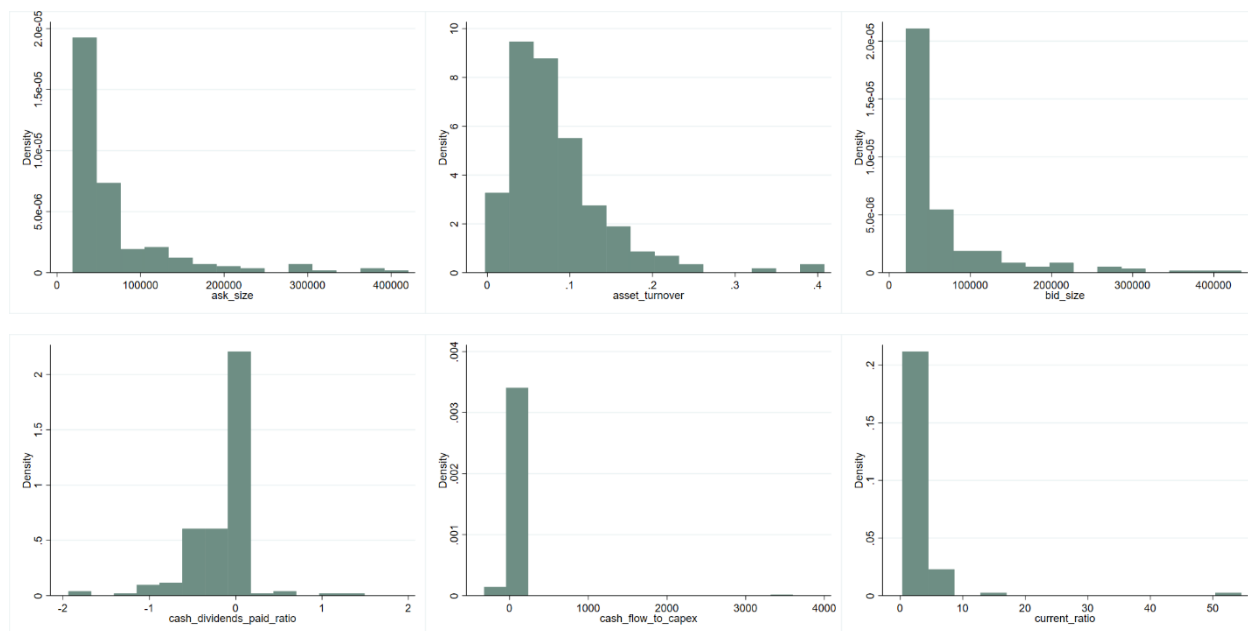
Stats	return_turn	return_turn	std_turn	target	upper_ratio	wap
Mean	.0130606	-1.531333	.1384046	-.0393863	1.333908	.9999772
SD	.0404022	20.42478	.0920344	.2109191	.2570515	.0001795
CV	3.093449	-13.33791	.664966	-5.355135	.1927056	.0001795

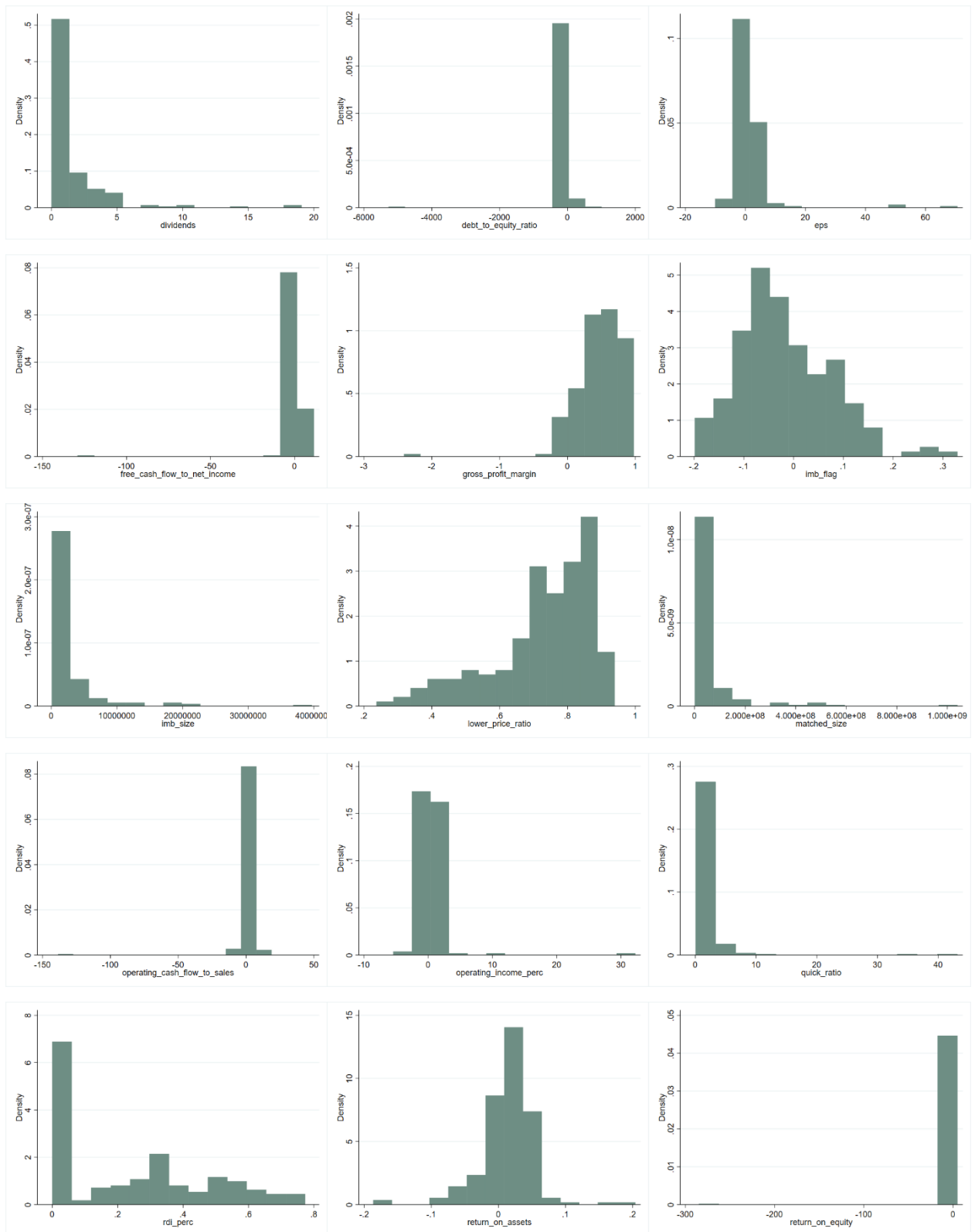
Źródło: opracowanie własne

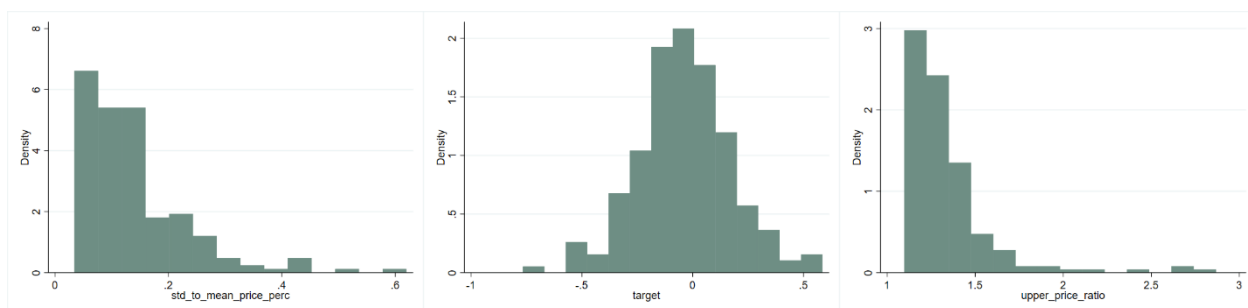
W rezultacie podjęto decyzję o usunięciu z analizowanego zbioru zmiennych o niskiej zdolności dyskryminacyjnej, których współczynnik zmienności był niższy od 0,1. Dotyczyło to zmiennych takich jak: *ask\_price*, *bid\_price*, *far\_price*, *near\_price*, *ref\_price* oraz *wap*. Takie zamiary sugerują niewielkie zróżnicowanie wewnątrz dniowe w czynnikach technicznych na przestrzeni całego roku.

Następnie analizie poddano rozkłady zaprezentowanych zmiennych.

Rysunek 1. Rozkłady zmiennych diagnostycznych







Źródło: opracowanie własne

W wyniku analizy graficznej zmiennych podjęto decyzję o przeprowadzeniu transformacji logarytmicznej dla następujących zmiennych: *ask\_size*, *at*, *bid\_size*, *dividends*, *imb\_size*, *matched\_size*, *q\_ratio*, *up\_ratio*, *lp\_ratio*, *stmpp*, *gpm*, *current\_ratio*. Niemniej jednak, po zastosowaniu procedury logarytmizacji, zdecydowano się na usunięcie zmiennych *ask\_size*, *bid\_size*, *imb\_size*, *matched\_size*, gdyż współczynniki zmienności dla tych zmiennych były niższe niż 0,1 po ponownym ich obliczeniu. Dla zmiennych *dte*, *fcftni*, *ocfts*, *oip*, *roe*, *eps* zastosowano winsoryzację poprzez usunięcie 2% największych i najmniejszych wartości zmiennych, zastępując je wartościami odpowiednio z 98 i 2 percentyla. Ostatecznie, zdecydowano się wykluczyć z analizy zmienną *rdip* z uwagi na jej praktycznie jednostajny rozkład.

W kolejnym etapie sprawdzono korelację między zmiennymi oraz na podstawie macierzy korelacji przeprowadzono procedury wyboru zmiennych diagnostycznych w oparciu o metodę Hellwiga.

Rysunek 2. Macierz korelacji zmiennych diagnostycznych oraz wyniki procedury wyboru zmiennych

	at_ls	cdp_ra~s	cftc_ws	curren~s	divide~s	dte_ra~s	eps_ws	fcftni~s	gpm_ls	imb_fl~s	lp_rat~s	ocfts_ws	oip_ws	q_rati~s	roa_s	roe_ws	stmpp_ls	target_s	up_rat~s
at_ls	1.0000																		
cdp_ratio_s	0.1004	1.0000																	
cftc_ws	0.0043	-0.0544	1.0000																
current_ra~s	0.3507	-0.0113	0.1897	1.0000															
dividends_ls	0.0190	-0.5005	-0.0626	-0.2721	1.0000														
dte_ratio_ws	-0.1182	0.1152	-0.0416	0.0140	-0.1341	1.0000													
eps_ws	-0.0206	-0.0534	-0.1501	-0.2038	0.2566	-0.0811	1.0000												
fcftni_ws	-0.0134	-0.1933	0.0511	-0.0630	0.0146	0.0034	-0.0261	1.0000											
gpm_ls	0.3806	-0.0568	0.1356	0.7828	-0.1670	-0.0494	-0.2455	-0.0410	1.0000										
imb_flag_s	-0.0793	0.1828	-0.0315	-0.0115	-0.3187	0.1532	-0.1479	-0.0470	-0.0191	1.0000									
lp_ratio_ls	0.0893	-0.2996	-0.0019	-0.0437	0.4559	-0.0347	0.1163	0.0444	0.0227	-0.5024	1.0000								
ocfts_ws	-0.0333	-0.3685	0.1198	-0.0720	0.1032	-0.0728	0.0658	0.8615	-0.0238	-0.0550	0.1230	1.0000							
oip_ws	0.1023	-0.1731	-0.1028	0.0801	0.3079	-0.1112	0.1775	0.0509	0.1866	-0.1253	0.2731	0.0933	1.0000						
q_ratio_ls	0.3070	-0.0064	0.1671	0.8666	-0.2368	-0.0010	-0.1588	-0.0556	0.6592	-0.0255	-0.0293	-0.0787	0.0663	1.0000					
roa_s	0.2140	-0.1252	-0.3498	0.0441	-0.2357	-0.0654	0.4966	0.0884	0.0170	-0.2175	0.3397	0.1228	0.3328	0.0184	1.0000				
roe_ws	-0.0207	0.0010	-0.0354	-0.0111	-0.0904	0.4559	-0.0095	0.0662	-0.0431	-0.0047	0.0803	0.0444	-0.1049	0.0076	-0.0064	1.0000			
stmpp_ls	-0.1292	0.3411	0.0143	0.0381	-0.5065	0.0347	-0.1304	-0.0072	-0.0209	0.4909	-0.8854	-0.0760	-0.2695	0.0041	-0.3279	-0.0639	1.0000		
target_s	-0.0179	0.0038	0.1001	-0.0188	0.1020	-0.0671	0.1078	-0.0832	-0.0871	-0.1721	-0.0372	-0.0643	-0.0133	0.0415	0.0025	0.0319	0.0388	1.0000	
up_ratio_ls	-0.1780	0.2310	0.0402	0.0483	-0.3964	0.1269	-0.2215	-0.0379	-0.0094	0.5743	-0.7687	-0.0953	-0.1952	-0.0094	-0.4143	-0.0839	0.8337	0.0561	1.0000

Zmienne reprezentanty	Typ zmiennej reprezentanta	Zmienne sat. dla zm.centralnych
up_ratio_ls	zmienna centralna	imb_flag_s lp_ratio_ls stmp_l
current_ratio_ls	zmienna centralna	gpm_ls q_ratio_ls
roa_s	zmienna izolowana	
ocfts_ws	zmienna centralna	fcftni_ws
dividends_ls	zmienna centralna	cdp_ratio_s
dte_ratio_ws	zmienna izolowana	
oip_ws	zmienna izolowana	
cftc_ws	zmienna izolowana	
target_s	zmienna izolowana	
at_ls	zmienna izolowana	
roe_ws	zmienna izolowana	
eps_ws	zmienna izolowana	

Źródło: opracowanie własne

Jednakże, z uwagi na stosunkowo niskie korelacje między zmiennymi, zdecydowano się na pozostawienie wszystkich wyselekcjonowanych zmiennych. Po zakończeniu transformacji i czyszczenia zbioru danych, analizowany zestaw składał się z 20 zmiennych diagnostycznych. Ostateczne statystyki opisujące te zmienne (przed standaryzacją) zostały przedstawione w poniższej tabeli.

Tabela 2. Statystyki opisowe zmiennych diagnostycznych

Mean	Std. dev.	Min	Max	Variable
-.1567236	.3524403	-1.937008	1.492806	cdp_ratio
-.0168669	.0932345	-.1984579	.3291898	imb_flag
.0130606	.0404022	-.1866324	.2040207	roa
-.0393863	.2109191	-.7658349	.5848858	stock_id
				target
-4.716932	4.950701	-9.21034	2.947072	dividends_l
-.3886046	2.488829	-9.21034	3.764964	q_ratio_l
-2.78642	.9769423	-9.21034	0	asset_turn~l
.2741796	.1576216	.0925572	1.053337	up_ratio_l
-.3431855	.2450707	-1.439081	-.0627675	lp_ratio_l
-2.152201	.5780384	-3.375293	-.4787134	stmp_l
-1.323795	2.243808	-9.21034	0	gpm_l
.1059837	2.267817	-9.21034	3.999077	current_ra~l
-8.353325	16.68043	-87.57853	32.79553	cftc_w
2.04289	23.05632	-62.15001	115.1033	dte_ratio_w
.8386299	2.024167	-6.545455	7.668383	fcftni_w
1.27245	2.321799	-6.181818	10.07379	ocfts_w
-.0171346	.4220323	-1.488047	1.028598	roe_w
.2597629	.5254288	-1.329294	2.266764	oip_w
1.362254	3.090275	-4.92	16.96	eps_w

Źródło: opracowanie własne



### III. Analiza empiryczna

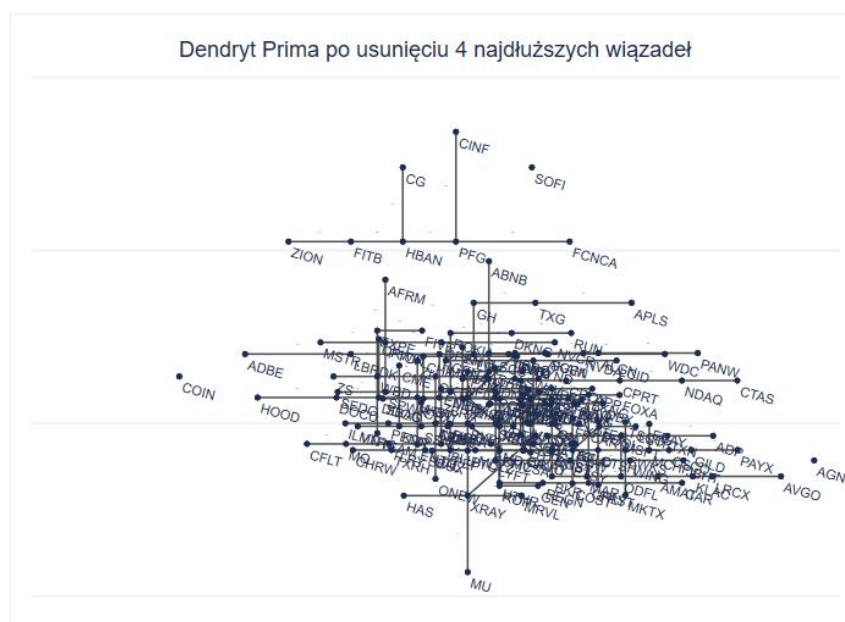
#### 3.1. Przebieg analizy

Część empiryczna rozpoczyna się od przedstawienia procedury analizy. W niniejszym badaniu wykorzystane zostaną różnorodne metody grupowania, takie jak: metoda podziału dendrytu, obszarowa metoda wrocławska, deglomeracyjna metoda najbliższej grupy oraz metody optymalizacji danego grupowania obiektów (metoda k-średnich, metoda Forgy-Janseya, oraz metoda Wisharta). W roli metody bazowej występują losowy podział grup oraz podział spółek według sektorów. Stanowią one solidny punkt odniesienia do oceny efektywności wybranych metod. Każda z metod jest przedstawiona w dwóch wersjach, gdyż dla każdej metody utworzono 5 oraz 10 grup spółek, co jest istotne dla zachowania porównywalności wyników. Jako główną miarę porównawczą wybrano miernik poprawności grupowania oparty na średniej arytmetycznej ze wszystkich mierników odległości. Dla najlepszej metody zostanie również opisana moc dyskryminacyjna zmiennych, a szczególna analiza zostanie skierowana na miary homogeniczności i heterogeniczności grup obiektów.

#### 3.2. Wyniki porównania użytych metod

W pierwszej kolejności dla pojedynczych akcji zbudowano dendryt, będący kluczowym elementem dla odpowiedniej metody. Następnie, po usunięciu 4 i 9 najdłuższych wiązań (co jest równoznaczne z utworzeniem 5 i 10 grup obiektów odpowiednio), skonstruowano dendrogram, umożliwiający klarowną identyfikację związku pomiędzy poszczególnymi spółkami.

Rysunek 3. Dendryt Prima po usunięciu 4 i 9 najdłuższych wiązań





Źródło: opracowanie własne

Z analizy graficznej powyższego wykresu jednoznacznie wynika, że istnieje wyraźny środek ciężkości, który stanowi centrum grupy zawierającej większość analizowanych spółek. Z perspektywy praktycznej, taki fakt może sugerować, że rozróżnienie akcji notowanych na giełdzie NASDAQ (przynajmniej według użytych w tym badaniu zmiennych diagnostycznych) jest niewielkie. Niemniej jednak powyższy rysunek ujawnia kluczową słabość metody podziału dendrytu, ponieważ wyodrębnione grupy często składają się zaledwie z jednego obiektu. Potwierdzeniem tego stanu rzeczy jest rozkład liczebności obiektów w grupach utworzonych za pomocą metody podziału dendrytu.

Tabela 3. Liczebności grup utworzonych za użyciem metody podziału dendrytu

dprima_gr10	Freq.	Percent	Cum.	dprima_gr5	Freq.	Percent	Cum.
1	183	93.37	93.37	1	185	94.39	94.39
2	1	0.51	93.88	2	1	0.51	94.90
3	1	0.51	94.39	3	1	0.51	95.41
4	1	0.51	94.90	4	1	0.51	95.92
5	1	0.51	95.41	5	8	4.08	100.00
6	1	0.51	95.92				
7	5	2.55	98.47				
8	1	0.51	98.98				
9	1	0.51	99.49				
10	1	0.51	100.00				
Total	196	100.00		Total	196	100.00	

Źródło: opracowanie własne

Z perspektywy analizy graficznej i aspektów ekonomicznych bardziej uzasadnionym podziałem dendrytu wydaje się zgrupowanie spółek położonych po lewej stronie od środka ciężkości

pierwszej grupy (tzn. utworzenie grupy niebieskiej), czyli akcje COIN, ADBE, HOOD, CFLT. Te spółki reprezentują przedsiębiorstwa znane głównie z jednego kluczowego produktu – oprogramowania (Coinbase i RobinHood Market, znane jako platformy do wymiany krypto walut i inwestowania bezprowizyjnego, Adobe, oraz Confluent, twórca Apache Kafka). Kolejną grupą stanowiłyby akcje położone powyżej środka ciężkości pierwszej grupy (grupa pomarańczowa): ZION, FITB, AFRM, HBAN, CG, PFG, ABNB, SOFI, CINF, FCNCA, firmy głównie świadczące usługi finansowe (banki, fundusze inwestycyjne) oraz platforma Airbnb. Akcje położone po prawej stronie od środka ciężkości pierwszej grupy tj.: APLS, WDC, PANW, NDAQ, CTAS, ADI, PAYX, AGNC, AVGO oraz z dołu (HAS, MU), mogłyby stanowić ostatnie odrębne grupy obserwacji odstających ze względu na użyte w badaniu zmienne diagnostyczne.

Kolejną zastosowaną metodą grupowania była metoda wrocławska. Podobnie jak w przypadku metody podziału dendrytu, ta metoda również cechowała się utworzeniem grup obiektów o niewielkiej liczebności. Dlatego już na danym etapie analizy można przypuszczać pewną bliską konkurencyjność obu metod grupowania.

Tabela 4. Liczebności grup utworzonych za użyciem metody wrocławskiej

mwroc_gr10	Freq.	Percent	Cum.				
1	180	91.84	91.84				
2	8	4.08	95.92				
3	1	0.51	96.43				
4	1	0.51	96.94				
5	1	0.51	97.45				
6	1	0.51	97.96				
7	1	0.51	98.47				
8	1	0.51	98.98				
9	1	0.51	99.49				
10	1	0.51	100.00				
Total	196	100.00					

mwroc_gr5	Freq.	Percent	Cum.
1	184	93.88	93.88
2	9	4.59	98.47
3	1	0.51	98.98
4	1	0.51	99.49
5	1	0.51	100.00
Total	196	100.00	

Źródło: opracowanie własne

Jeśli chodzi o same grupy, można zauważyć podobieństwo do zgrupowań utworzonych za pomocą metody podziału dendrytu. Grupa 2 składa się z akcji, które przypominają skład grupy 2 w metodzie podziału dendrytu, obejmując takie spółki, jak PFG, HBAN, SOFI, CG, AGNC, ZION, FCNCA, FITB, ADBE.

Następnie w analizie użyto metody deglomeracyjnej. W przypadku tej metody utworzone poszczególne klastry charakteryzowały się znacznie większą liczebnością przy ustawieniu pięciu grup obiektów. Niemniej jednak, przy ustawieniu 10 grup, wyniki były bardziej zbliżone do rezultatów uzyskanych za pomocą poprzednich metod, co sugeruje pewną konsystencję w grupowaniu przedsiębiorstw notowanych na NASDAQ w kontekście większej liczby klastrów.

Tabela 5. Liczebności grup utworzonych za użyciem metody deglomeracyjnej

degl_gr10	Freq.	Percent	Cum.	degl_gr5	Freq.	Percent	Cum.
1	1	0.51	0.51	1	62	31.63	31.63
2	1	0.51	1.02	2	20	10.20	41.84
3	2	1.02	2.04	3	102	52.04	93.88
4	4	2.04	4.08	4	4	2.04	95.92
5	1	0.51	4.59	5	8	4.08	100.00
6	3	1.53	6.12				
7	6	3.06	9.18				
8	96	48.98	58.16				
9	20	10.20	68.37				
10	62	31.63	100.00				
Total	196	100.00		Total	196	100.00	

Źródło: opracowanie własne

W kolejnym etapie uzyskano wyniki grupowań za pomocą metod optymalizacji zgrupowań: metody k-średnich, metody Forgy-Janseya oraz metody Wisharta.

Tabela 6. Liczebności grup utworzonych za użyciem metod optymalizacji grupowania

kmeans_gr10	Freq.	Percent	Cum.	kmeans_gr5	Freq.	Percent	Cum.
1	33	16.84	16.84	1	91	46.43	46.43
2	4	2.04	18.88	2	4	2.04	48.47
3	3	1.53	20.41	3	10	5.10	53.57
4	24	12.24	32.65	4	84	42.86	96.43
5	6	3.06	35.71	5	7	3.57	100.00
6	7	3.57	39.29				
7	43	21.94	61.22				
8	14	7.14	68.37				
9	7	3.57	71.94				
10	55	28.06	100.00				
Total	196	100.00		Total	196	100.00	

forjan_gr10	Freq.	Percent	Cum.	forjan_gr5	Freq.	Percent	Cum.
1	35	17.86	17.86	1	46	23.47	23.47
2	4	2.04	19.90	2	4	2.04	25.51
3	2	1.02	20.92	3	10	5.10	30.61
4	22	11.22	32.14	4	64	32.65	63.27
5	5	2.55	34.69	5	72	36.73	100.00
6	8	4.08	38.78				
7	47	23.98	62.76				
8	14	7.14	69.90				
9	8	4.08	73.98				
10	51	26.02	100.00				
Total	196	100.00		Total	196	100.00	

wishart_gr10	Freq.	Percent	Cum.	wishart_gr5	Freq.	Percent	Cum.
1	39	19.90	19.90	1	62	31.63	31.63
2	3	1.53	21.43	2	3	1.53	33.16
3	2	1.02	22.45	3	6	3.06	36.22
4	21	10.71	33.16	4	58	29.59	65.82
5	4	2.04	35.20	5	67	34.18	100.00
6	6	3.06	38.27				
7	42	21.43	59.69				
8	12	6.12	65.82				
9	8	4.08	69.90				
10	59	30.10	100.00				
Total	196	100.00		Total	196	100.00	

Źródło: opracowanie własne

Wyniki grupowań uzyskane za pomocą tych metod stanowiły znaczne odstępstwo w stosunku do pozostałych technik. Niemniej jednak, rezultaty osiągnięte pomiędzy metodami optymalizacji, zarówno ze względu na rozkłady liczebności, jak i średnich wartości zmiennych diagnostycznych w grupach, różniły się niewiele. Jedyne różnice udało się ujawnić podczas kolejnej analizy porównawczej wszystkich technik grupowania.

Ostatnia kategoria metod grupowania obejmowała techniki naiwne: zgrupowanie losowe oraz zgrupowanie według sektorów.

Tabela 7. Liczebności grup utworzonych za użyciem metod naiwnych

random_gr10	Freq.	Percent	Cum.	random_gr5	Freq.	Percent	Cum.
1	11	5.61	5.61	1	18	9.18	9.18
2	21	10.71	16.33	2	46	23.47	32.65
3	26	13.27	29.59	3	55	28.06	60.71
4	20	10.20	39.80	4	55	28.06	88.78
5	24	12.24	52.04	5	22	11.22	100.00
6	26	13.27	65.31				
7	23	11.73	77.04				
8	22	11.22	88.27				
9	13	6.63	94.90				
10	10	5.10	100.00				
Total	196	100.00		Total	196	100.00	

sector_gr10	Freq.	Percent	Cum.	sector_gr5	Freq.	Percent	Cum.
1	11	5.61	5.61	1	23	11.73	11.73
2	9	4.59	10.20	2	31	15.82	27.55
3	31	15.82	26.02	3	27	13.78	41.33
4	8	4.08	30.10	4	39	19.90	61.22
5	23	11.73	41.84	5	76	38.78	100.00
6	31	15.82	57.65				
7	15	7.65	65.31				
8	18	9.18	74.49				
9	38	19.39	93.88				
10	12	6.12	100.00				
Total	196	100.00		Total	196	100.00	

Źródło: opracowanie własne

Na tym punkcie analizy widoczna jest zauważalna różnica między metodami bazowymi a pozostałymi technikami grupowania. Sugeruje to, że opisane wcześniej metody są bardziej efektywne niż podział losowy i sektorowy, i prowadzą do bardziej sensownego zgrupowania.

Nowy etap analizy rozpoczyna się od zbadania średnich wartości zmiennych diagnostycznych w utworzonych grupach obiektów.

Rysunek 4. Średnie wartości zmiennych diagnostycznych w grupach



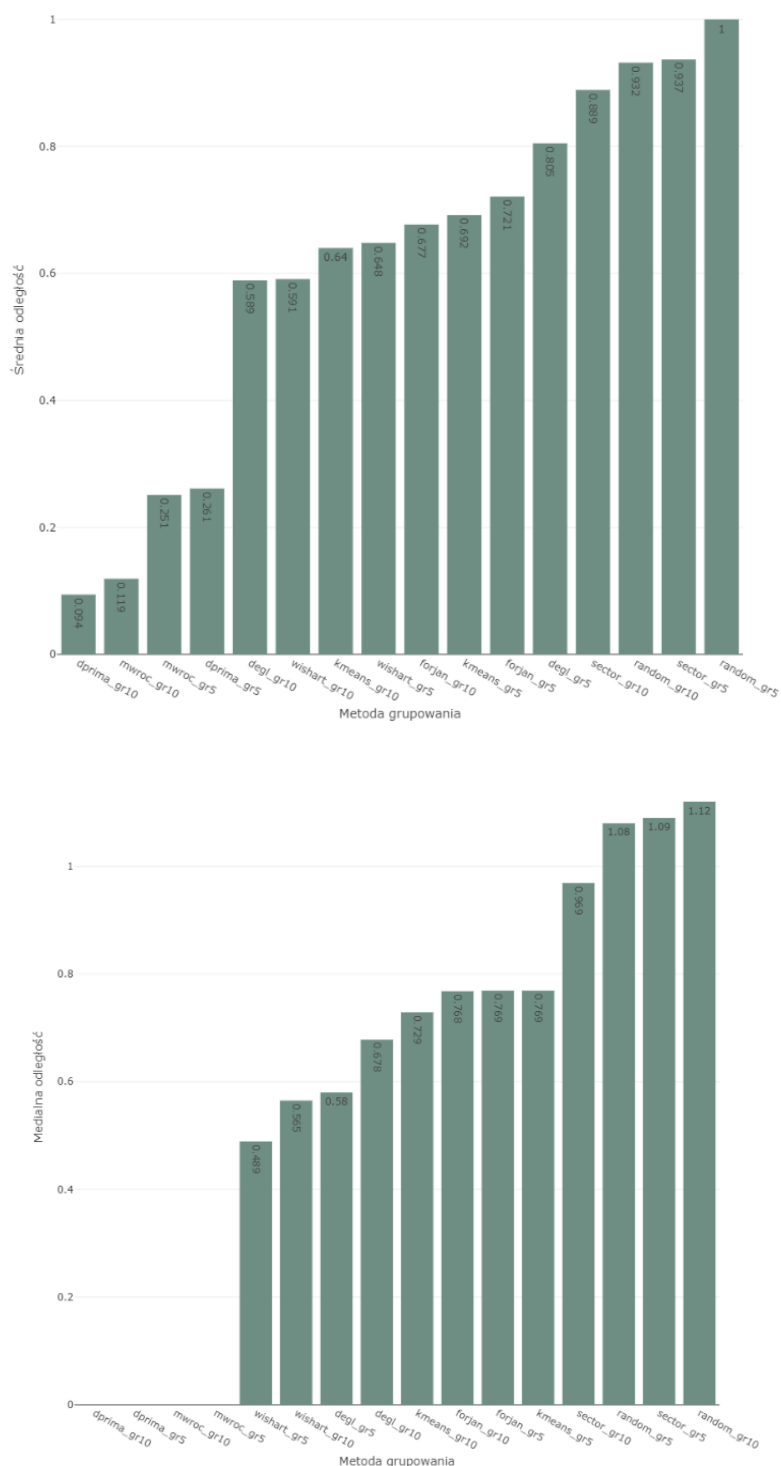




Źródło: opracowanie własne

W przypadku utworzonych 10 grup akcji obserwuje się znaczące zróżnicowanie średnich wartości zmiennych, zwłaszcza przy użyciu metody podziału dendrytu, metody wrocławskiej, metody deglomeracyjnej oraz metody k-średnich. W tych przypadkach odmienności międzygrupowe w średnich wartościach były bardziej wyraźne w porównaniu do pozostałych metod. Z kolei, przy 5

grupach spółek, każda z metod charakteryzowała się istotnym zróżnicowaniem średnich wartości zmiennych diagnostycznych. Niemniej jednak najbardziej widoczna dysproporcja między zastosowanymi metodami jest zauważalna w kontekście współczynników poprawności grupowania. Rysunek 5. Współczynniki poprawności grupowania



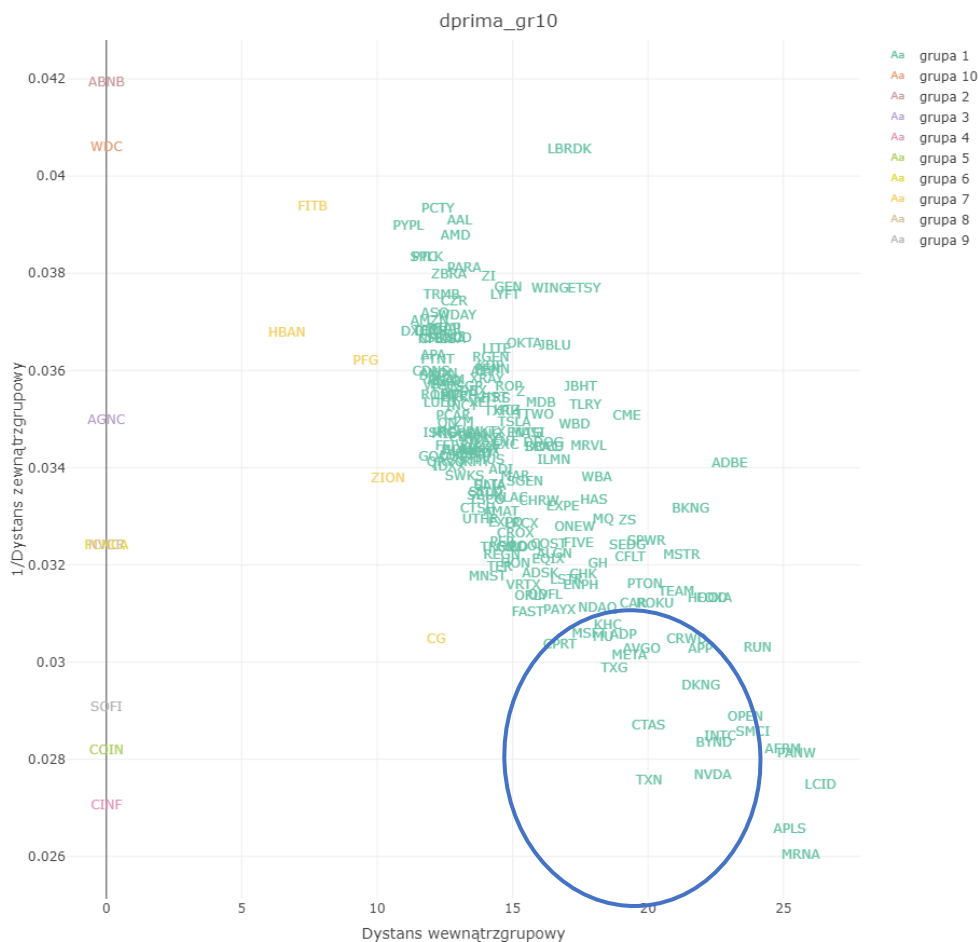
Źródło: opracowanie własne



Przedstawiony wykres pokazuje średnie oraz medialne odległości wewnątrzgrupowe i międzygrupowe, posortowane rosnąco. Warto zauważyć, że niższa wartość wskazuje na lepszy balans jednorodności wewnątrzgrupowej i odróżnienia międzygrupowego. Najbardziej skuteczną metodą grupowania akcji notowanych na giełdzie NASDAQ wydaje się być metoda podziału dendrytu dla 10 grup, charakteryzująca się wartością współczynnika poprawności grupowania równą 0,094. Na drugim miejscu plasuje się metoda wrocławska, co potwierdza wcześniejsze obserwacje dotyczące podobieństwa obu metod. Warto jednak zaznaczyć, że jakość grupowania za pomocą każdej metody statystycznej okazała się znacznie lepsza od podejścia podziału według sektorów oraz podziału losowego. Z kolei wyniki klasteryzacji za pomocą metod optymalizacji grupowania różniły się między sobą niewiele.

Dla najlepszej metody grupowania został również utworzony wykres przedstawiający odległość wewnątrz- i międzygrupową dla każdej akcji. Pozwala to łatwo zidentyfikować potencjalne obserwacje odstające, które mogłyby być poddane kolejnemu procesowi grupowania.

Rysunek 6. Średnia odległość wewnątrz- i zewnątrz grupowa dla metody podziału dendrytu



Źródło: opracowanie własne

W ten sposób można zauważyć, że wśród grupy 1 występuje przynajmniej jeden przypadek nietypowy (spółka LBRDK). Co więcej, dolny ogon grupy 1 można było poddać dodatkowemu procesowi grupowania w celu zmniejszenia odległości wewnątrzgrupowej (obszar zaznaczony na niebiesko). Dodatkowo, z analizy graficznej wynika, że metoda podziału dendrytu stara się w pierwszej kolejności zidentyfikować obserwacje odstające. Dlatego też metoda ta może być szczególnie przydatna dla inwestorów, którzy chcą skupić się na najbardziej nietypowych przypadkach, co naturalnie może prowadzić do osiągania zysków przewyższających rynek lub unikania pewnych znaczących strat.

Ostatnia tabela prezentuje wyniki pomiaru mocy dyskryminacyjnej dla zmiennych diagnostycznych, testowanych za pomocą statystyki F-Snedecora.

Tabela 8. Moc dyskryminacyjna zmiennych diagnostycznych

Zmienna	BSS	df 1	WSS	df 2	F	Pr > F
at_ls	55.58	9	136.53	186	8.41	0.0000
cdp_ratio_s	24.45	9	173.19	186	2.92	0.0030
cftc_ws	62.45	9	135.30	186	9.54	0.0000
current_ratio_ls	161.06	9	19.90	186	167.29	0.0000
dividends_ls	13.39	9	180.93	186	1.53	0.1404
dte_ratio_ws	50.36	9	139.83	186	7.44	0.0000
eps_ws	31.87	9	140.20	186	4.70	0.0000
fcftni_ws	40.67	9	157.28	186	5.34	0.0000
gpm_ls	118.93	9	66.37	186	37.03	0.0000
imb_flag_s	27.94	9	168.21	186	3.43	0.0006
lp_ratio_ls	31.76	9	162.15	186	4.05	0.0001
ocfts_ws	36.68	9	161.29	186	4.70	0.0000
oip_ws	15.33	9	181.62	186	1.74	0.0816
q_ratio_ls	121.29	9	64.06	186	39.13	0.0000
roa_s	26.51	9	169.77	186	3.23	0.0012
roe_ws	28.59	9	156.68	186	3.77	0.0002
stmp_ls	17.37	9	178.92	186	2.01	0.0407
target_s	8.28	9	189.11	186	0.90	0.5222
up_ratio_ls	36.15	9	161.32	186	4.63	0.0000

*Źródło: opracowanie własne*

Jak wynika z tabeli, prawie wszystkie używane zmienne miały istotny wkład w rozróżnienie utworzonych 10 grup spółek. Najlepszą mocą dyskryminacyjną charakteryzowały się zmienne mierzące wysokość wypłaconych dywidend w ujęciu procentowym w ciągu roku oraz średnia wartość syntetyczna ceny akcji w ostatnich 5 minutach przed zamknięciem targów. W ten sposób można stwierdzić, że wpływ czynników związanych z handlem wśród dziennym na rozróżnianie spółek

notowanych na giełdzie NASDAQ jest nieistotny, gdyż zmienna *target\_s* była ostatnim czynnikiem tego typu.

## Zakończenie

Na podstawie analizy empirycznej stwierdzono, że najlepszą metodą do grupowania cen akcji notowanych na giełdzie papierów wartościowych NASDAQ jest metoda podziału dendrytu z 10 grupami. Miara poprawności grupowania wykazała, że wyniki grupowania za pomocą metody podziału dendrogramu były 10 razy lepsze od podejścia polegającego na podziale według sektorów i grupowania losowego. Co więcej, zaletą tej metody jest łatwa interpretacja graficzna uzyskanych wyników. Dodatkowo udało się podkreślić nieistotność czynników związanych z handlem wśród dziennym w kształtowaniu relacji zachodzących pomiędzy akcjami. Również zaobserwowano, że mimo pozornej różnorodności sektorów i samych notowanych przedsiębiorstw, w większości przypadków mają one podobne zachowania długoterminowe, poza kilkoma obserwacjami odstającymi.

Należy zaznaczyć, że badanie to posiada szereg ograniczeń. Po pierwsze, dotyczyło jedynie wyników analizy danych za 2023 rok. W celu stworzenia bardziej kompleksowej i odpornej analizy konieczne byłoby rozszerzenie horyzontu czasowego, co mogłoby przynieść ciekawe perspektywy zrozumienia ewolucji rynku. Po drugie, ze względu na różnorodność użytych metod i pragnienie zachowania porównywalności wyników, nie można było dostosować liczbę grup obiektów  $k$  za pomocą automatycznych metod optymalizacyjnych. Po trzecie, kryterium poprawności grupowania użyty w niniejszej pracy stanowi tylko jedno z wielu możliwych kryterium i może nie w pełni odzwierciedlać całość związków i procesów ekonomicznych zachodzących na giełdzie NASDAQ. Pomimo tych ograniczeń, niniejsze badanie posiada potencjał do praktycznego zastosowania w kontekście takim, jak optymalizacja portfeli inwestycyjnych na podstawie utworzonych grup i obliczenie zwrotów rocznych w porównaniu do indeksu giełdowego. Warto również zaznaczyć, że w dalszych badaniach można rozważyć uwzględnienie dodatkowych czynników wpływających na dynamikę cen akcji, które biorą pod uwagę szerszy kontekst ekonomiczny i społeczny.

## Bibliografia

1. Al-Augby S., Majewski S., Majewska A., Nermend K., *A Comparison Of K-Means and Fuzzy C-Means Clustering Methods for A Sample of Gulf Cooperation Council Stock Markets*, Folia Oeconomica Stetinensia, Wyd. 14 Nr. 2, 2014, s. 19-36.
2. Andreassen P. B., *On the social psychology of the stock market. Aggregate attributional effects and the regressiveness of prediction*, Journal of Personality and Social Psychology Wyd. 53 Nr. 3, 1987, s. 490–496.
3. Arthur W. B., *Silicon Valley locational clusters: when do increasing returns imply monopoly?*, Mathematical Social Sciences Wyd. 19 Nr. 3, 1990, s. 235-251.
4. Azzopardi P.V., *Behavioral Technical Analysis: An introduction to behavioral finance and its role in technical analysis*, Harriman House, Nowy York 2010.
5. Babu S., Geethanjali N., Satyanarayana B., *Clustering Approach to Stock Market*, Int. J. Advanced Networking and Applications Wyd. 3 Nr. 4, 2012, s. 1281-1291.
6. Bini B.S., Tessy M., *Clustering and Regression Techniques for Stock Prediction*, Procedia Technology Wyd. 24, 2016, s. 1248-1255.
7. Calinski R. H., *A dendrite method for cluster analysis*, Communications in Statistics Wyd. 3, 1974, s. 1–27.
8. Chaudhuri D., Chaudhuri B. B., *A novel multiseed nonhierarchical data clustering technique*, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) Wyd. 27 Nr. 5, 1997, s. 871-887.
9. Chiu C. Y., Chen Y. F., Kuo I. T., Ku H. C., *An intelligent market segmentation system using k-means and particle swarm optimization*, Expert Systems with Applications Wyd. 36 Nr. 3, 2009, s. 4558-4565.
10. Dattatray D., Gandhmal P., Kumar K., *Systematic analysis and review of stock market prediction techniques*, Computer Science Review Wyd. 34, 2019, s. 1000-1019.
11. Fritz H., García-Escudero L. A., Mayo-Isar A., *A fast algorithm for robust constrained clustering*, Computational Statistics & Data Analysis Wyd. 61, 2013, s. 124-136.
12. Grossman S. J., Stiglitz J. E., *On the Impossibility of Informationally Efficient Markets*, American Economic Review, 1980, s. 393–408.
13. Kumar P., Wasan S.K., *Comparative Analysis of k-mean Based Algorithms*, International Journal of Computer Science and Network Security Wyd. 10 Nr. 4, 2010, s. 314–318.

14. Kuo R. J., Wang H. S., Hu T. I., Chou S. H., *Application of ant K-means on clustering analysis*, *Computers & Mathematics with Applications*, Wyd. 50 Nr. 10–12, 2005, s. 1709-1724.
15. Kyoung-jae K., Hyunchul A., *A recommender system using GA K-means clustering in an online shopping market*, *Expert Systems with Applications* Wyd. 34 Nr. 2, 2008, s. 1200-1209.
16. Murphy J. J., *Technical analysis of the financial markets: a comprehensive guide to trading methods and applications*, New York Institute of Finance, Nowy York 1999, s. 1-33.
17. Nanda S. R., Mahanty B., Tiwari M. K., *Clustering Indian stock market data for portfolio management*, *Expert Systems with Applications* Wyd. 37 Nr. 12, 2010, s. 8793-8798.
18. Oti E. U., Olusola M. O., Eze F. C., Enogwe S. U., *Comprehensive Review of K-Means Clustering Algorithms*, *International Journal of Advances in Scientific Research and Engineering* Wyd. 7 Nr. 8, 2021, s. 64-69.
19. Panek T., Zwierzchowski J., *Statystyczne metody wielomiarowej analizy porównawczej*, Oficyna Wydawnicza SGH, Warszawa 2013, s. 33-185.
20. Wu D., Wang X., Wu S., *Construction of stock portfolios based on k-means clustering of continuous trend features*, *Knowledge-Based Systems* Wyd. 252, 2022, s. 1093-1058.
21. Xindong W., Kumar V., *The top ten algorithms in Data Mining*, *Knowl Inf Syst* Wyd. 14, 2008, s. 21–36.
22. Zheng F., Chaoshin C., *Research on Prediction and Recommendation of Financial Stocks Based on K-means Clustering Algorithm Optimization*, *Journal of Computational Methods in Sciences and Engineering*, Wyd. 21 Nr. 5, 2021, s. 1081 – 1089.
23. <https://www.kaggle.com/competitions/optiver-trading-at-the-close/overview>, dostęp 20.01.2024.
24. <https://www.nasdaq.com/>, dostęp 20.01.2024.
25. <https://finance.yahoo.com/>, dostęp 20.01.2024.

## Spis rysunków

Rysunek 1. Rozkłady zmiennych diagnostycznych .....	12
Rysunek 2. Macierz korelacji zmiennych diagnostycznych oraz wyniki procedury wyboru zmiennych.....	14
Rysunek 3. Dendryt Prima po usunięciu 4 i 9 najdłuższych wiązań .....	16
Rysunek 4. Średnie wartości zmiennych diagnostycznych w grupach .....	21
Rysunek 5. Współczynniki poprawności grupowania.....	23
Rysunek 6. Średnia odległość wewnątrz- i zewnątrz grupowa dla metody podziału dendrytu .....	24

## Spis tabeli

Tabela 1. Współczynniki zmienności zmiennych diagnostycznych .....	12
Tabela 2. Statystyki opisowe zmiennych diagnostycznych.....	15
Tabela 3. Liczebności grup utworzonych za użyciem metody podziału dendrytu.....	17
Tabela 4. Liczebności grup utworzonych za użyciem metody wrocławskiej .....	18
Tabela 5. Liczebności grup utworzonych za użyciem metody deglomeracyjnej .....	19
Tabela 6. Liczebności grup utworzonych za użyciem metod optymalizacji grupowania .....	19
Tabela 7. Liczebności grup utworzonych za użyciem metod naiwnych .....	20
Tabela 8. Moc dyskryminacyjna zmiennych diagnostycznych .....	25