ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»
УДК 519.868

Отчет об исследовательском проекте

на тему Кластеризация цен акций энергетических компаний

Выполнен студентом:
Группы #БПМИ2210 Евтушенко Олег Владимирович
ФИО студента

Проверен руководителем проекта:

Касьянова Ксения Алексеевна (магистр)

стажер - исследователь
Должность

Международная лаборатория стохастического анализа и его приложений

Москва 2024

ывапрол

1.Аннотация

Данная работа посвящена исследованию методов межвременной кластеризации цен акций энергетических компаний. Характер развития рынка акций носит динамический характер и подвержен влиянию многих факторов. Эффективный анализ и классификация ценовых движений с помощью машинного обучения, включая рассматриваемую методику кластеризации, может стать одним из ключевых инструментов принятия решений для инвесторов и финансовых аналитиков. В работе рассматриваются такие методы кластеризации как – иерархическая кластеризация, метод k-средних, моделирование смеси Гауссовых распределений. Также будет рассмотрен для сравнительного анализа метод сглаживания данных- ядерная оценка плотности в рамках метода среднего сдвига и будет рассмотрена возможность интеграции метода динамического искривления времени из статьи Khadoudja Ghanem для кластеризации цен акций энергетических компаний.

Ключевые слова – машинное обучение, кластеризация, ядерная оценка плотности, распределение Гаусса, динамическое искривление времени.

This study is dedicated to methods of intertemporal clustering of stock prices for energy companies. Influenced by various factors, the stock market exhibits the dynamic nature. Thus, effective analysis and classification of price movements through machine learning, including the considered clustering methodology, can be regarded as a crucial decision-making tool for investors and financial analysts. The study delves into clustering methods such as hierarchical clustering, k-means clustering, modeling of Gaussian mixture distributions. The study will also exploit the kernel density estimation in MeanShift method and explore the potential integration of the dynamic time warping method from Khadoudja Ghanem's article for clustering in terms of stock prices of energy companies.

Keywords: machine learning, clustering, kernel density estimation, Gaussian distribution, dynamic time warping.

2.Введение

2.1 Описание предметной области

В современном мире финансовых рынков акции энергетических компаний играют важную роль, обеспечивая не только устойчивость, но и продвижение национальных и мировых экономик. Энергетический сектор, будучи ключевым столпом мировой промышленности, подвергается постоянным воздействиям различных внешних факторов, что влечет за собой колебание цен акций

Рынок акций в свою очередь представляет из себя сложную динамическую систему со множеством взаимосвязей, где тысячи факторов могут влиять на ценовые движения. Эффективный анализ и классификация таких движений с помощью машинного обучения может стать одним из ключевых инструментов принятия решений для инвесторов и финансовых аналитиков.

2.2 Постановка задачи

Целью данного исследования является сравнительный анализ различных методов кластеризации (иерархическая кластеризация, метод k-средних, ядерная оценка плотности, моделирование смеси Гауссовых распределений) для выявления временных паттернов и структур в ценовых движениях акций энергетических компаний.
Каждый из выбранных методов будет рассмотрен с точки зрения их применимости, эффективности и устойчивости к динамике рынка. Кроме того, возможно будет рассмотрена возможность потенциальной интеграции метода динамического искривления времени, предложенного Khadoudja Ghanem, для повышения эффективности построения моделей.

3. Обзор литературы

3.1 Иерархическая кластеризация

Иерархическая кластеризация — это метод анализа данных, который используется для группировки объектов или точек данных в иерархическую структуру. Данный метод создает дерево (или дендрограмму), которое иллюстрирует отношения между различными кластерами.
Выделяют 2 класса методов Иерархической кластеризации:
Агломеративные методы - новые кластеры создаются путем объединения более мелких кластеров - дерево создается от листьев к стволу;
Дивизивные или *дивизионные методы* - новые кластеры создаются путем деления более крупных кластеров на более мелкие и, таким образом, дерево создается от ствола к листьям

Авторы статьи James Ming Chen, Mobeen Ur Rehman, Xuan Vinh Vo (2021) использовали данный подход для построения кластеров, с помощью которых визуализировали и интерпретировали логарифмические доходы и условную волатильности в торговле товарами. Данный подход я собираюсь применить для анализа цен акций энергетических компаний, что станет одной из отправных точек для сравнительного анализа с другими методами.

3.2 Метод k-средних

Метод k-средних (k-Means Clustering) –распространенный и довольно мощный алгоритм Обучения без учителя (Unsupervised Learning), который группирует похожие элементы в k кластеров.

В общем случае метод делится на 3 основных этапа. На исходном этапе выбирается значение k оно будет определять количество кластеров, которые получатся в готовой модели. Затем инициализируются центроиды или же разделительные линии. Дальше выбирается группа и ищется среднее значение расстояния между точками. Процесс продолжается до тех пор, пока не будут перебраны все возможные сочетания пар дистанцированных точек и не будут уточнены границы кластеров.

В своей статье Fernandez-Avil´es et al. (2020) авторы предложили использовать данный подход для построения кластеров, связанных с торговлей товарами. Однако, их модель была построена без применения DTW (dynamic time warping) и сами авторы с своей статье указали на возможность применения других алгоритмов кластеризации или использования других метрик для дальнейшего исследования.

3.3 ядерная оценка плотности

Ядерная оценка плотности – не является конкретно видом кластеризации, однако представляет собой метод сглаживания данных, который используется для формирования выводов о распределении данных, основываясь на ограниченных выборках информации.

В своей статье Adriano Z. Zambom and Ronaldo Dias (2013) применили данный метод оценки таких параметров как ориентировочная плотность вознаграждения генерального директора и оценки экспорта оценки товаров и услуг, однако их исследование не затронуло анализ энергетического сектора. К тому же в статье Chen J (2021) авторы также указали на возможность применения метода ядерной оценки плотности для дальнейшего исследования рынка товаров и сектора энергетики.

3.4 Моделирование смеси Гауссовых распределений

Гауссова смесь распределений - статистическая модель для представления нормально распределенных субпопуляций внутри общей популяции.

Гауссова смесь распределений параметризируется двумя типами значений — смесь *весов компонентов* и *средних компонентов* или *ковариаций* (для *многомерного* случая). Если количество компонентов известно, техника, чаще всего используемая для оценки параметров смеси распределений — EM-алгоритм.

В статье Gaussian mixture and financial return (2008) исследователи сделали вывод, данный подход может быть использован для определения распределения доходности активов и выявления различных настроений на фондовом рынке. Также в статье Chen J (2021) авторы сослались на то, что их исследование может быть продолжено и в качестве одной из альтернатив указали использование моделирование смеси Гауссовых распределений.

4. Данные и метрики

4.1 Пул компаний

В рамках данного исследования было принято решение изучать цены акций энергетических компаний в рамках одной страны (России). Данный подход позволяет исключить/сгладить влияние различных макроэкономических факторов (политическая стабильность, изменения валютного курса, и т.д.), которые не только могут оказывать влияние на цены акций, но и усложнять анализ в целом. Помимо этого, проведение такого анализа внутри одной страны также облегчает сравнение результатов и выявление особых тенденций, специфичных для исследуемого сектора.

Компании, которые были выбраны для проведения исследования:

Название сферы	Компании (тикеры)
Нефтегазовая промышленность	"Газпром" (GAZP.ME) "Лукойл" (LKOH.ME) "Роснефть" (ROSN.ME) "Транснефть" (TRNFP.ME) "Газпром нефть" (SIBN.ME) "Татнефть" (TATN.ME) "Башнефть" (BANE.ME) "Сургутнефтегаз" (SNGS.ME) "НОВАТЭК" (NVTK.ME)
Химическая промышленность	Фос Агро (PHOR.ME) Акрон (AKRN.ME)
Энергетика	"РусГидро" (HYDR.ME) "Мосэнерго" (MSNG.ME) "Интер РАО ЕЭС" (IRAO.ME) "Трансгаз Москва" (TRMK.ME)
Металлургия	"Русал" (RUAL.ME) "Северсталь" (CHMF.ME)
Транспорт (включено в эксперимент. целях)	"Аэрофлот" (AFLT.ME)
Электроэнергетика	"Россети" (RSTI.ME) +

Таблица 1 «Пул компаний»

4.2 Разведочный анализ данных (EDA)

Разведочный анализ данных — анализ основных свойств данных, который проводится в целях нахождения общих закономерностей, распределений, возможных отклонений и аномалий. Обычно с использованием инструментов визуализации.

Непосредственно наши данные:

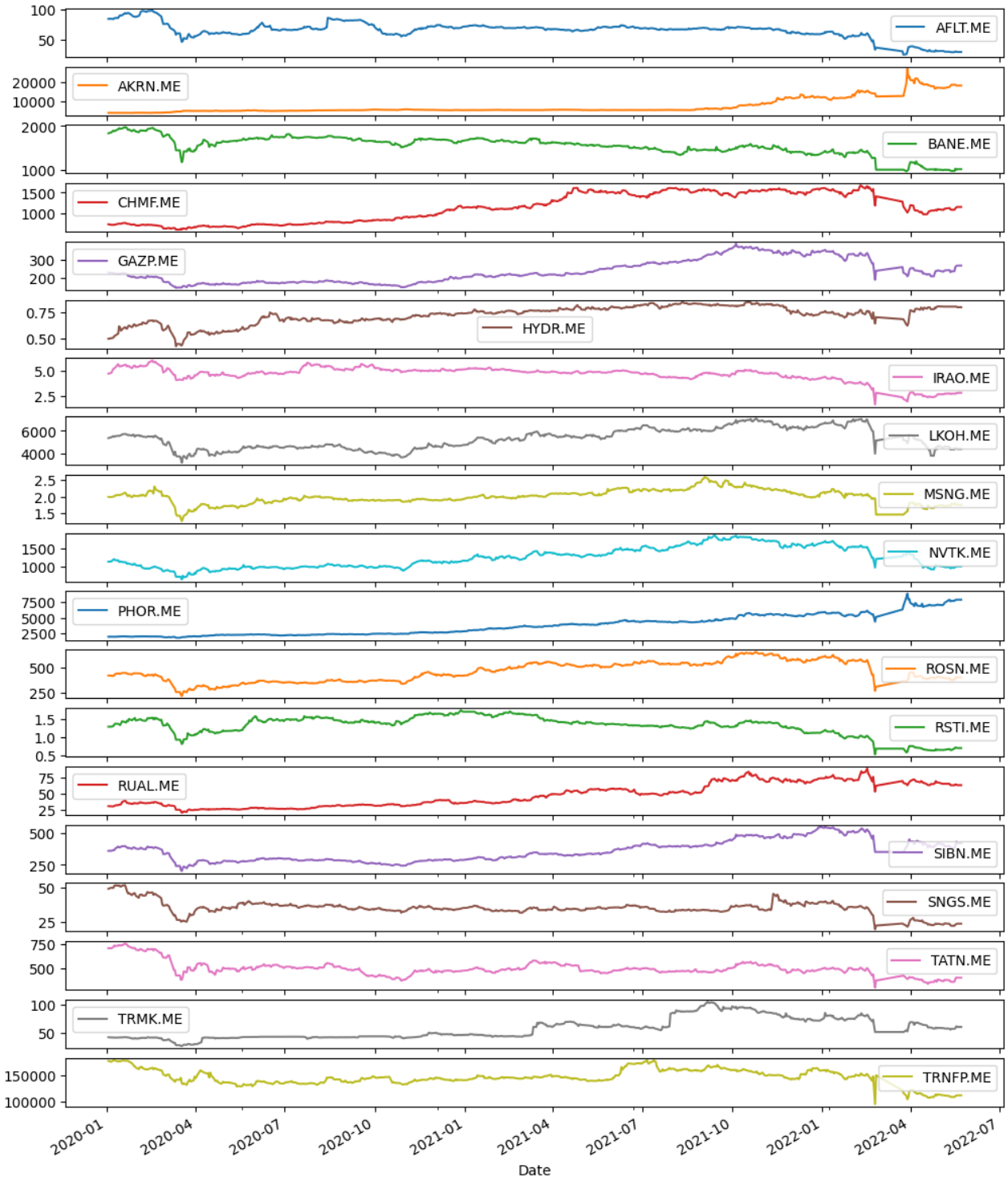


График 1 Цены акций за период с 2020-01-01 по 2022-05-24

Построим тепловую карту, которая в общем случае помогает определить степень взаимосвязи между переменными

Данная цветовая схема 'coolwarm' наглядного отображает положительную (более теплые цвета) и отрицательную (более холодные цвета) корреляцию между ценами исследуемых акций.

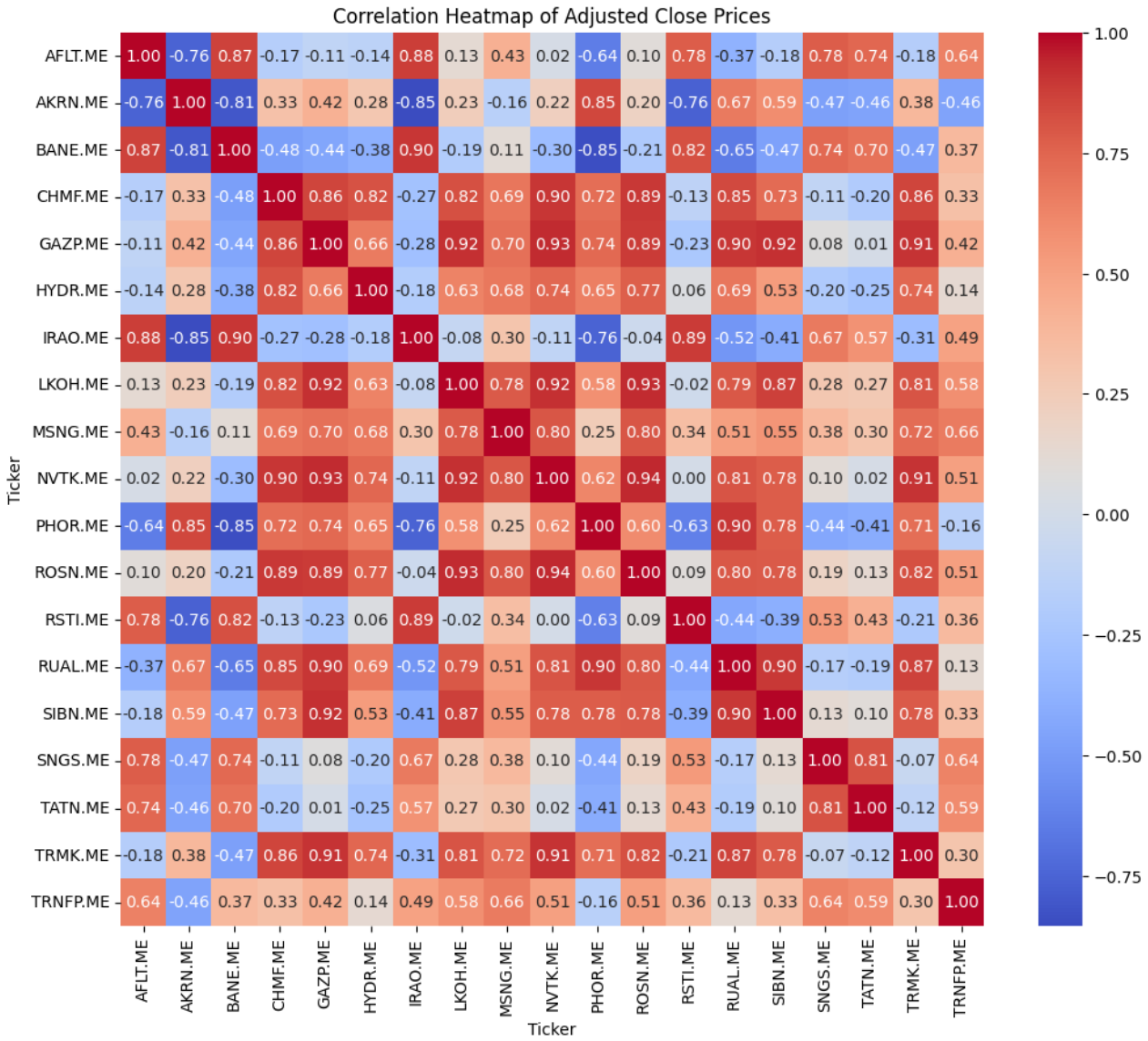


График 2 Heatmap для исследуемых акций

Краткие выводы, которые можно сделать на основе таблицы:

1) Наибольшей корреляцией с другими ценными бумагами обладает NVTK.ME ("НОВАТЭК")

2) Наименьшей корреляцией с другими ценными бумагами обладает BANE.ME ("Башнефть")

(Подробные результаты сравнения можно увидеть в разделе Приложения, таблица 3)

Также в рамках EDA вычислим сводные статистики и меры центральной тенденции:

Ticker	AFLT.ME	AKRN.ME	BANE.ME	CHMF.ME	GAZP.ME	HYDR.ME	IRAO.ME	LKOH.ME	MSNG.ME	NVTK.ME	PHOR.ME	ROSN.ME	RSTI.ME
count	581.000000	581.000000	581.000000	581.000000	581.000000	581.000000	581.000000	581.000000	581.000000	581.000000	581.000000	581.000000	581.000000
mean	66.088538	7296.700903	1556.844121	1130.498467	234.407323	0.719984	4.682505	5317.036558	2.000815	1260.846078	3753.518823	457.857552	1.326463
std	12.715795	4033.510477	204.797047	351.936841	65.160630	0.085645	0.718344	914.910313	0.213222	292.567390	1590.560841	100.636926	0.252994
min	25.100000	3974.327148	960.000000	597.933228	144.446060	0.425818	1.692500	3201.289307	1.250943	647.009705	1789.186768	214.374283	0.521600
25%	60.779999	5363.347656	1446.500000	760.834778	176.251999	0.672491	4.446500	4555.483887	1.891152	1000.072144	2325.052246	366.793762	1.237600
50%	67.379997	5534.049805	1588.953735	1112.961548	216.968979	0.737400	4.848398	5367.318359	1.999244	1217.379028	3552.812256	445.376343	1.381291
75%	70.680000	6476.320312	1695.500000	1507.800049	284.233887	0.790508	5.085293	6084.623535	2.140500	1524.637573	4784.912109	540.159058	1.481324
max	98.148720	27010.000000	1974.411621	1676.000000	386.149994	0.843000	6.031559	7069.947266	2.599500	1876.724243	8908.000000	653.599976	1.742105

RUAL.ME	SIBN.ME	SNGS.ME	TATN.ME	TRMK.ME	TRNFP.ME
581.000000	581.000000	581.000000	581.000000	581.000000	581.000000
46.990947	355.903764	35.251811	497.573325	57.974617	146525.861231
17.516321	80.972196	4.958600	72.597637	18.529309	14929.300319
20.410000	202.090668	19.525000	302.500000	27.219236	94350.000000
31.719999	286.634735	33.595001	467.549408	43.074112	138817.531250
40.465000	336.190521	35.142914	488.399994	50.064209	145473.781250
64.019997	404.377167	36.975346	517.450562	70.300003	157600.000000
88.955002	545.049988	52.112766	755.468689	105.752495	179306.046875

Таблица 4 (сводные статистики и меры центральной тенденции)

Краткие выводы: есть переменные с разными диапазонами значений, что может затруднять интерпретацию результатов. Для некоторых методов, которые чувствительны к данной ситуаций придется также нормализовать имеющиеся данные.

4.3 Метрики

Одним из самых главных аспектов является выбор метрик, с помощью которых построенные модели будут сравниваться между собой. В рамках данной работы были выбраны следующие метрики:

1) Силуэт (Silhouette) - мера, которая оценивает, насколько каждый объект данных похож на свой собственный кластер (это называется "сплоченность"), по сравнению с другими кластерами. Значение индекса Силуэта находится варьируется от -1 до +1. Высокое значение указывает на то, что объект хорошо подходит к своему кластеру и плохо соответствует соседним кластерам. Низкое или отрицательное значение для многих точек может указывать на то, что в кластеризации либо слишком много, либо слишком мало кластеров. Классификация по средним значениями: «сильное» – более 0,7, «разумное» - более 0,5, «слабое» - более 0,25.

Однако при работе с данными высокой размерности достижение высоких значений индекса Силуэта может быть затруднено из-за проклятия размерности, когда расстояния между точками становятся более схожими. Несмотря на это, силуэт является одним из самых распространенных метрик в рамках задачи кластеризации.

2) Индекс Калински-Харабаша (Calinski-Harabasz) – данный показатель оценивает, насколько хорошо данные были разбиты на кластеры, не опираясь на какие-либо заранее известные ответы или метки. Индекс оценивает качество кластеризации исключительно на основе структуры данных и результатов самой кластеризации.

$$CH = \frac{BCSS/(k-1)}{WCSS/(n-k)}$$

Определяется данный индекс как , где в свою очередь BCSS - это взвешенная сумма квадратов евклидовых расстояний между каждым центроидом кластера и общим центроидом данных. А WCSS в свою очередь определяется как сумма квадратов евклидовых расстояний между точками данных и их соответствующими центроидами кластера.

3) Индекс Дэвиса –Булдина (DBI) – индекс, который также оценивает качество кластеризации, учитывая как компактность кластеров, так и возможность их разделения. Низкое значение этого индекса указывает на лучшую кластеризацию.

На основе вышеперечисленных метрик и будем строить наш дальнейший анализ. Ввиду того, что каждый из методов кластеризации уже был подробно описан в разделе «Обзор литературы», перейдем непосредственно к построению наших моделей и расчетам наших метрик

5. Построение моделей

5.1 Иерархическая кластеризация

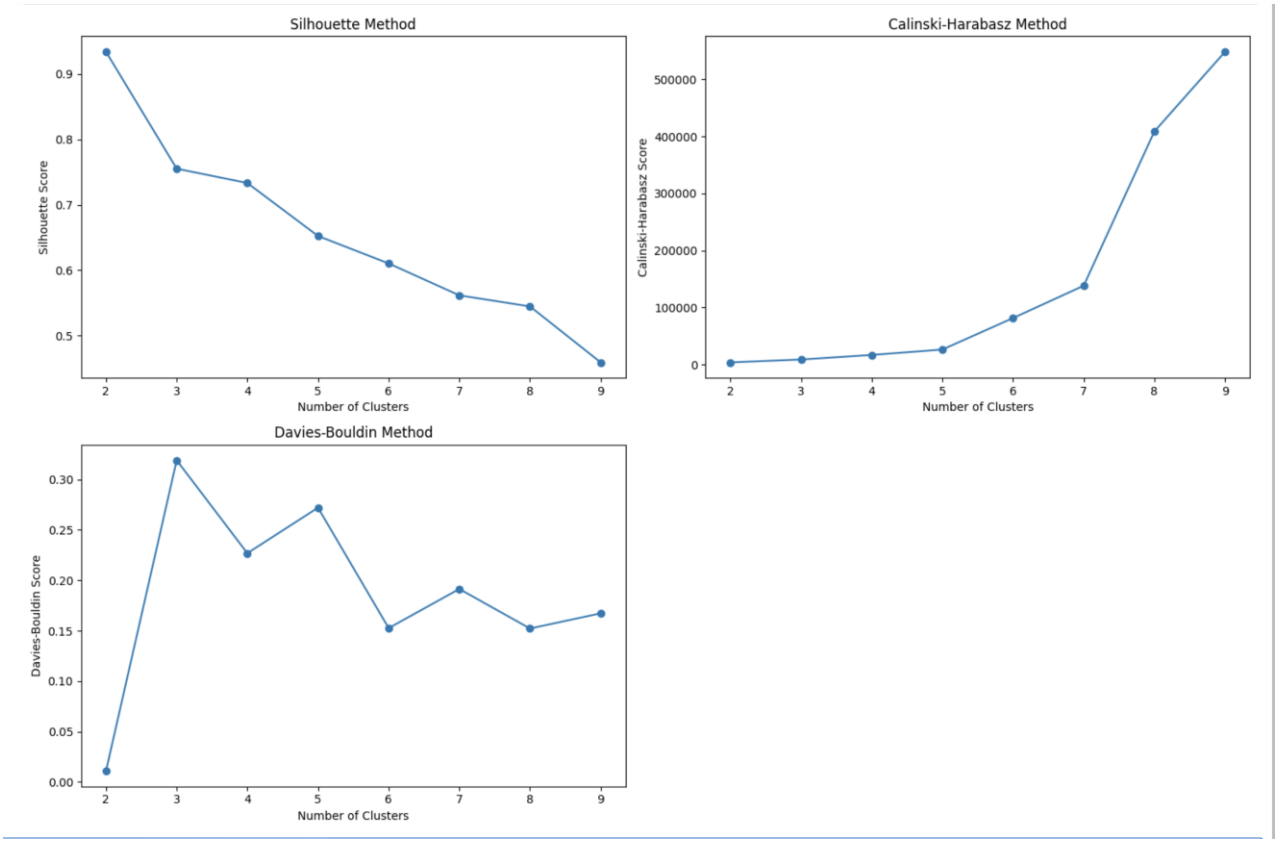


График 3 (Подбор количества кластеров для иерархической кластеризации)

С учетом всех выбранных метрик для данной модели получим наилучшим значением количества кластеров n=8 (берется средневзвешенное по всем метрикам). Соответственно визуализация и расчеты

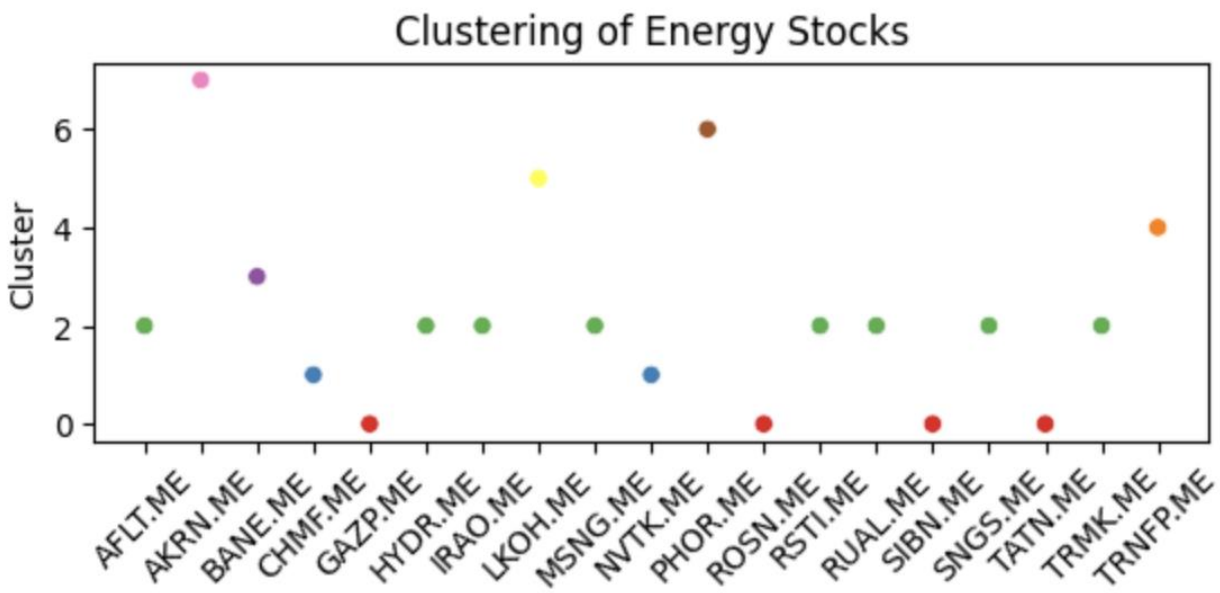


График 4 (иерархическая кластеризация)

Также можно изобразить в форме дендограммы

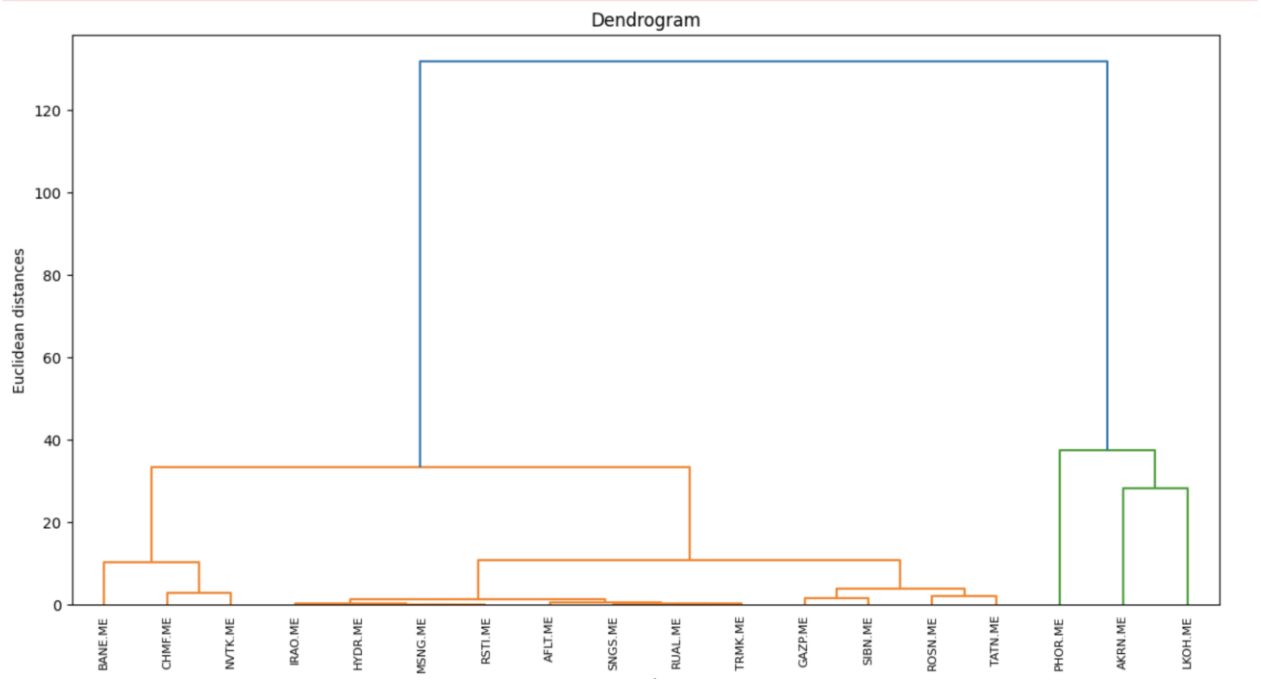


График 5 (Иерархическая кластеризация – дендограмма Ward’s method)

Значения метрик для анализа

Silhouette Score: 0.5250822225414438
Calinski-Harabasz Score: 1836.011071465924
Davies-Bouldin Score: 0.26522704180934653

5.2 KMeans

Аналогично иерархической кластеризации начнем с поиска количества кластеров

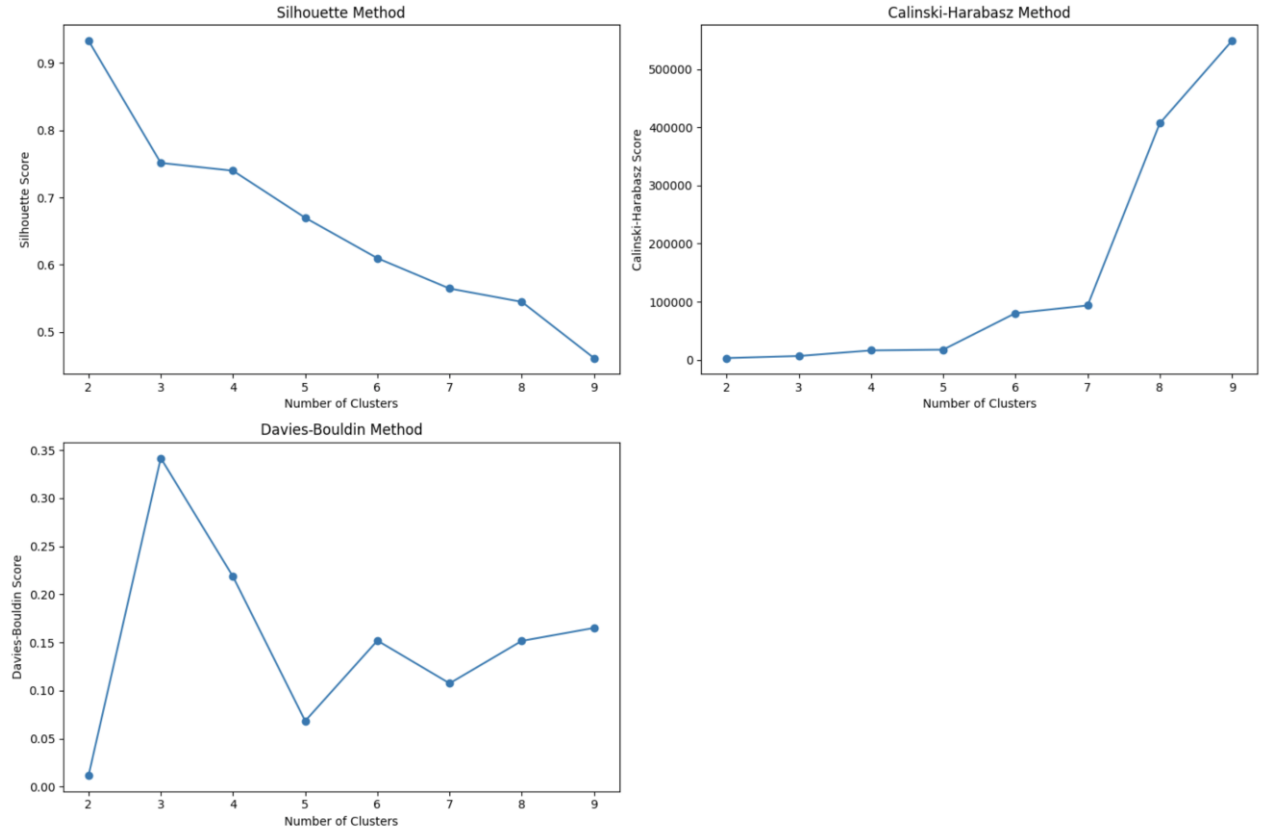


График 6 (Подбор числа кластеров для Kmeans)

Уже на данном этапе можем наблюдать различия графиков и характера изменения значений по сравнению с иерархической кластеризацией. В данном случае наилучшим значением количества кластеров уже будет n=7. Построим модель и вычислим характеристики.

Для визуализации модели воспользуемся методами PCA и t-SNE. Считается, что это улучшает результаты кластеризации за счёт уменьшения шума, а также используется для визуализации. Данные методы также используются для понижения размерности данных, что облегчает процесс визуализации

(графики построены ниже)

Непосредственно расчеты метрик (PCA)

Silhouette Score: 0.43767274989518007
Calinski-Harabasz Score: 622.117594531074
Davies-Bouldin Score: 0.1979527795982421

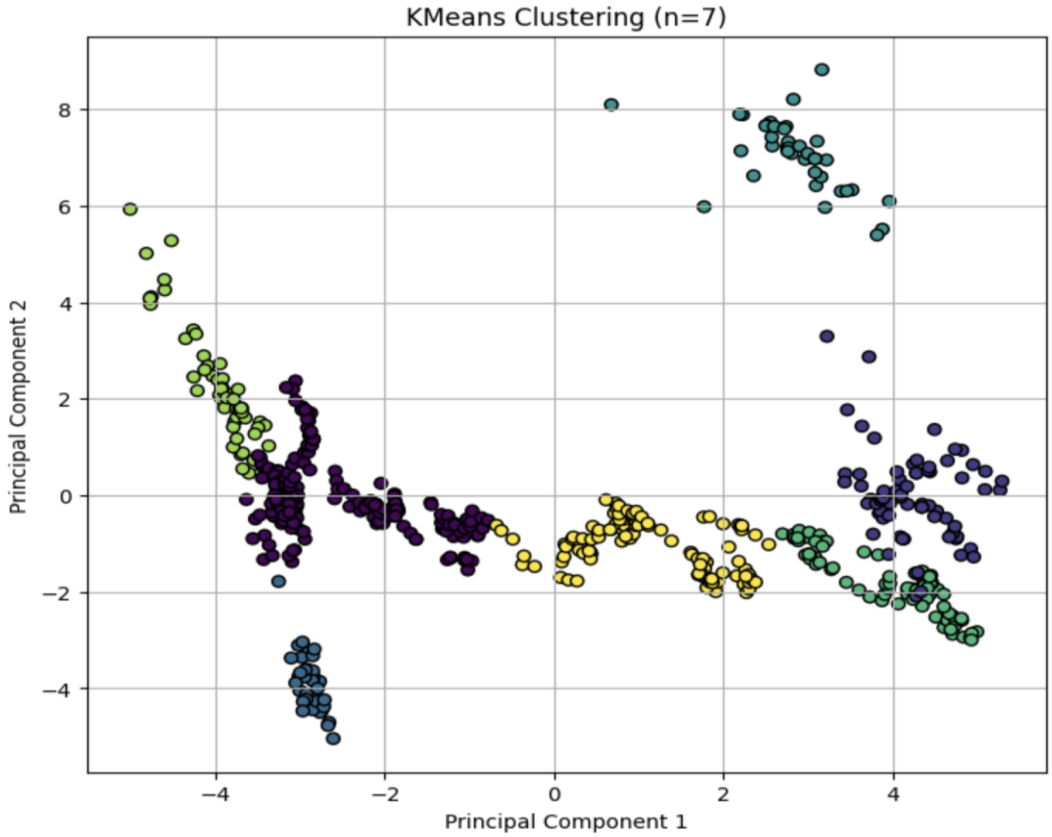
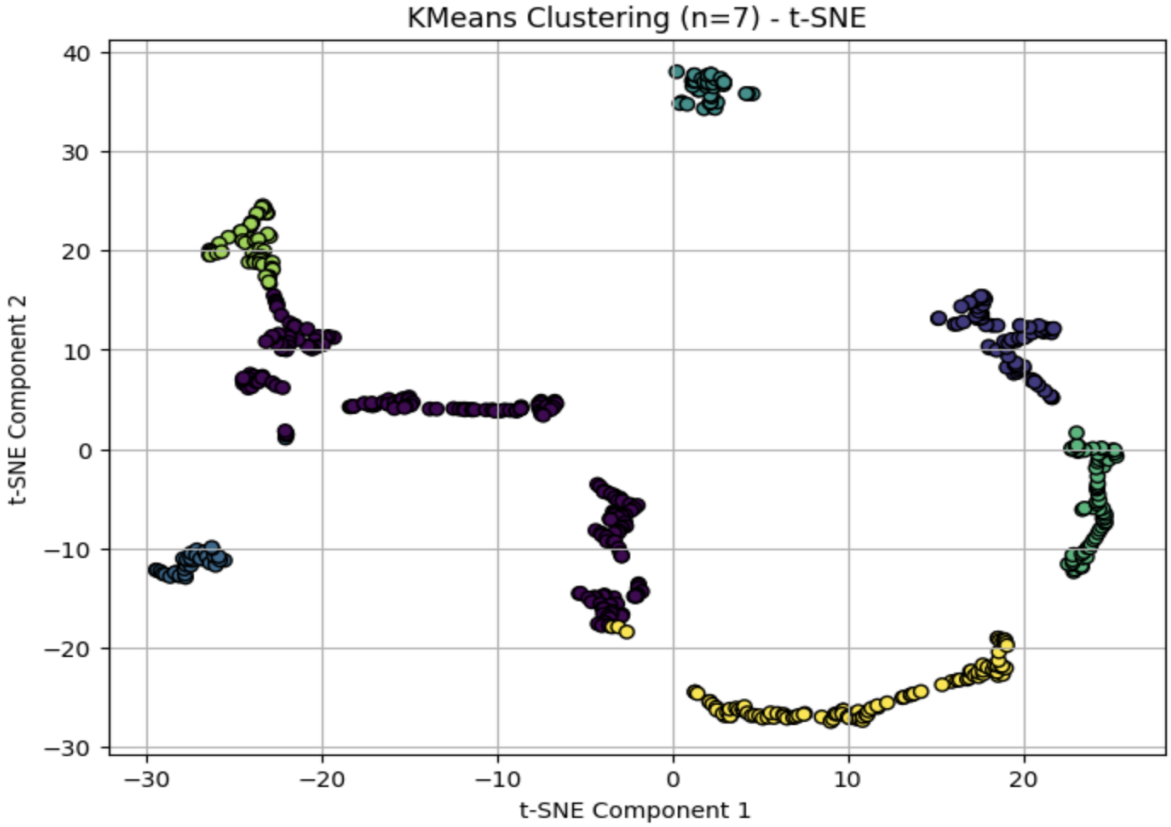


График 7 (визуализация Kmeans с PCA)




```
Silhouette Score: 0.43767274989518007
Calinski-Harabasz Score: 622.117594531074
Davies-Bouldin Score: 0.7979527795982421
```

График 8 (Визуализация Kmeans с t-SNE)

Однако, для большей наглядности изобразим в трехмерном пространстве, отобразив относительно наших акций.

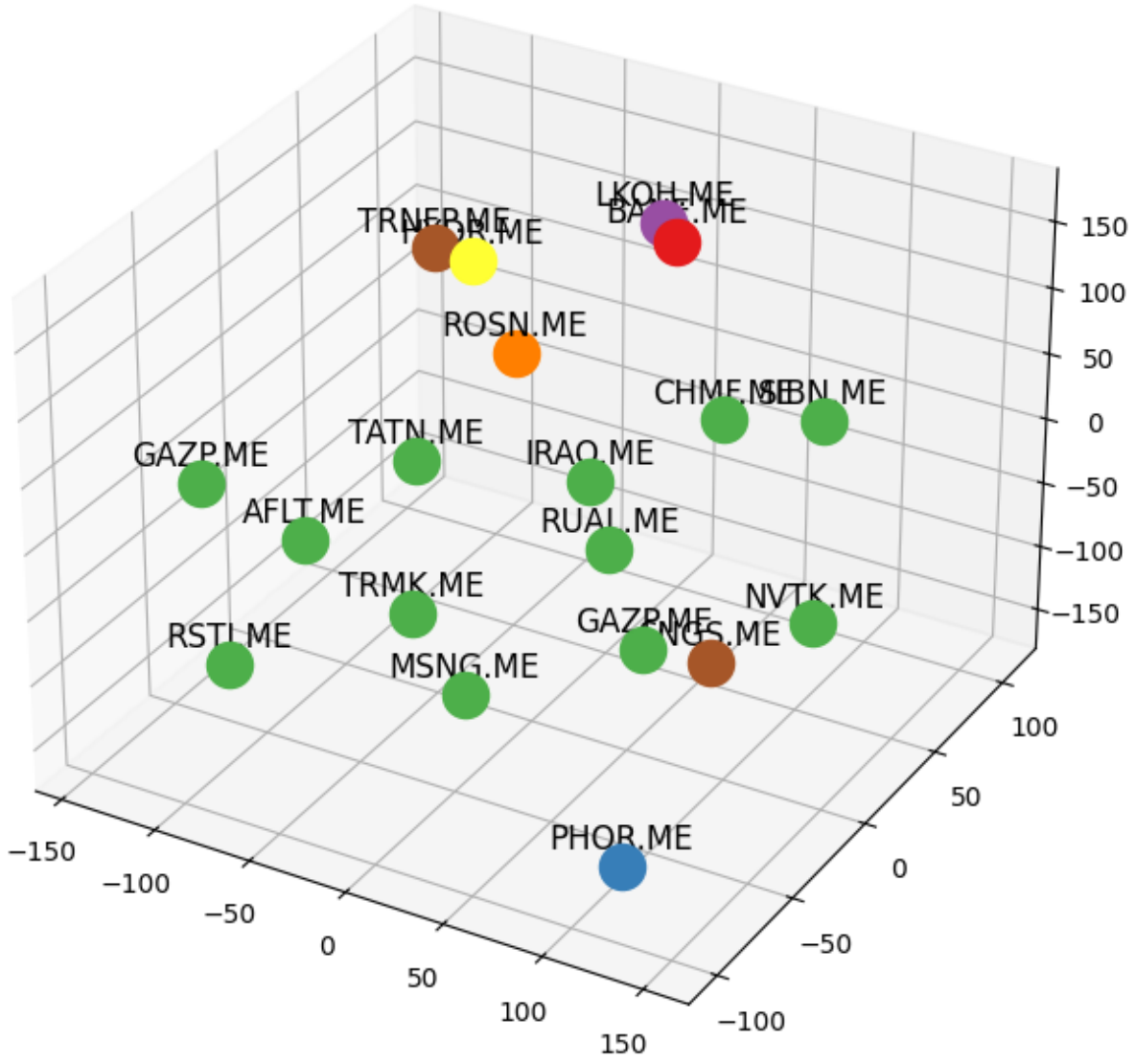


График 9 (Kmeans в трехмерном пространстве, MDS)

Одинаковым цветом выделены те акции, которые входят соответственно в один и тот же кластер. Теперь можно наглядно оценить взаимосвязь между различными акциями и выявить возможные паттерны или тренды на рынке при необходимости.

5.3 Смеси Гауссовских распределений

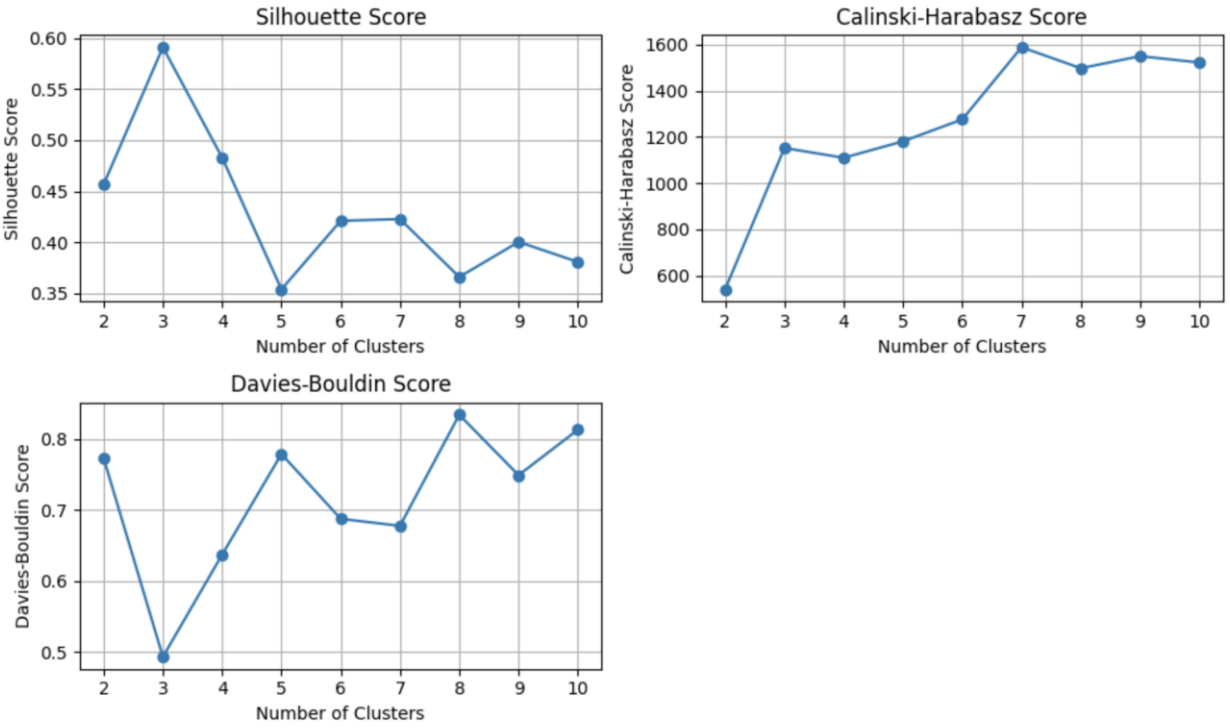
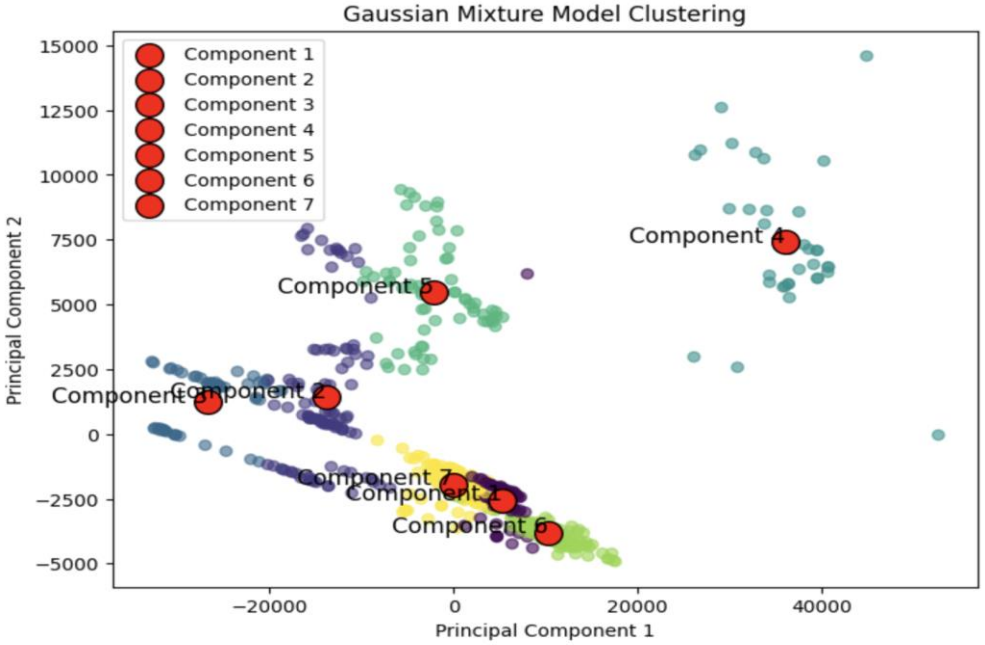


График 10 (подбор числа кластеров для Гауссовской смеси)

Можно наблюдать довольно интересную ситуацию с 2 метриками – DBI и Silhouette. При n=3 одна функция достигает максимума, в то время как другая достигает минимума, аналогично при n=8, что свидетельствует о том, что в кластеризации нельзя полагаться только на 1 метрику, нужно проводить либо комплексный анализ, либо смотреть на цель исследования и уже в зависимости от нее принимать решение насчет использования той или иной метрики. В нашем случае оптимумом будет n=7.




```
Silhouette Score: 0.42277928605211335
Calinski-Harabasz Score: 1588.1568211692138
Davies-Bouldin Score: 0.6779239063204158
```

График 11 (Кластеризация на основе Гауссовских смесей)

5.4 Оценка плотности ядра (kernel density estimation), метод среднего сдвига MeanShift.

Данный способ используется для представления данных в виде гладкой кривой. По сути, берется каждая точка данных и над ней ставится некое "ядро" (как правило, это гауссовское распределение), после чего все эти ядра складываются вместе. Таким образом, получается кривая, которая помогает понять, как данные распределены по значениям.

Также стоит отметить, что оценка плотности ядра является первым шагом в кластеризации методом среднего сдвига. Поэтому в рамках данного исследования сравним метрики, полученные путем применения среднего сдвига с остальными моделями, которые уже были рассмотрены. Метод среднего сдвига носит непараметрический характер (нас не интересует количество кластеров), поэтому сразу приведем результаты вычисления метрик:

Silhouette Score	Calinski-Harabasz Score	Davies-Bouldin Score
0.630816	205.38971	0.321072

Таблица 5 (результаты метода среднего сдвига)

6. Выводы и результаты

В первую очередь хочется отметить, что не существует универсального метода кластеризации, как и метрики для оценки модели, которые всегда были бы лучше остальных. Выбор подходящей техники кластеризации и метрики зависит как от характера данных, целях проводимого исследования, так и порой от предпочтений исследователя. В рамках нашей задачи, при средневзвешенном подходе к выбору количества кластеров мы получили, что:

1) Для Silhouette Score наилучшим методом оказался – MeanShift с оценкой плотности ядра = 0,630816
2) Для Calinski-Harabasz наилучшим методом оказалась - Иерархическая кластеризация (Ward’s method) = 1836.011071465924
3) Для Davies-Boudlin-Index наилучшим методом оказался – Kmeans с PCA = 0.1979527795982421

Таким образом, каждый из исследуемых методов проявил себя в той или иной метрике, каждый из подходов обладает своими сильными и слабыми сторонами. Нельзя однозначно судить что в рамках данной задачи какой-то из рассмотренных методов является наилучшим, к тому же сами метрики отвечают за различные параметры и свойства кластеров.

В заключение, хотелось бы отметить возможные пути продолжения данного исследования – метод динамического искривления времени не был по итогу рассмотрен. Также возможно изучение влияния дополнительных факторов, таких как объемы торгов или новостные события и есть возможность учета дополнительных метрик, которые также можно было бы попробовать сравнить.

Литература

[1]. James Ming Chen, Mobeen Ur Rehman, Xuan Vinh Vo. Clustering commodity markets in space and time: Clarifying returns, volatility, and trading regimes through unsupervised machine learning (2021)

[2]. Ming-Heng Zhan, Gaussian Mixture Model to Detect Random Walks in Capital Markets

[3]. Fern'andez-Avil´es, G., Montero, J.-M., Sanchis-Marco, L., 2020. Extreme downside risk comovement during distress periods: a multidimensional scaling approach. Eur. J.Finance 26, 1207–1237

[4]. Adriano Z. Zambom and Ronaldo Dias, A Review of Kernel Density Estimation with Applications to Econometrics

[5]. Carlos Cuevas Covarrubias, Anahuac University, Mexico. Jorge Rosales Contreras, ITESM, Mexico, Gaussian Mixture and Financial Return March (2006)

[6]. Khadoudja Ghanem, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.2, March 2013

[7]. Capo, M., Perez, A., Lozano, J.A., 2017. An efficient approximation to the k-means clustering for massive data. Knowl. Base Syst. 117, 56–69

[8]. Forster, R., 2006. Document clustering in large German corpora using natural language processing (Doctoral dissertation, University of Zürich)

Приложения

	Ticker Most.	Most Correlated	Correlation	Least Correlated	Correlation
0	AFLT.ME	IRAO.ME	0.882231	AKRN.ME	-0.755586
1	AKRN.ME	PHOR.ME	0.850718	IRAO.ME	-0.853636
2	BANE.ME	IRAO.ME	0.898683	PHOR.ME	-0.845528
3	CHMF.ME	NVTK.ME	0.903585	BANE.ME	0.479550
4	GAZP.ME	NVTK.ME	0.928139	BANE.ME	-0.444917
5	HYDR.ME	CHMF.ME	0.816376	BANE.ME	-0.384744
6	IRAO.ME	BANE.ME	0.898683	AKRN.ME	-0.853636
7	LKOH.ME	ROSN.ME	0.933917	BANE.ME	-0.191321
8	MSNG.ME	NVTK.ME	0.804347	AKRN.ME	-0.158208
9	NVTK.ME	ROSN.ME	0.939176	BANE.ME	-0.304990
10	PHOR.ME	RUAL.ME	0.902484	BANE.ME	-0.845528
11	ROSN.ME	NVTK.ME	0.939176	BANE.ME	-0.213991
12	RSTI.ME	IRAO.ME	0.889752	AKRN.ME	-0.764497
13	RUAL.ME	GAZP.ME	0.902682	BANE.ME	-0.649532
14	SIBN.ME	GAZP.ME	0.919641	BANE.ME	-0.466928
15	SNGS.ME	TATN.ME	0.812492	AKRN.ME	-0.474340
16	TATN.ME	SNGS.ME	0.812492	AKRN.ME	-0.459931
17	TRMK.ME	NVTK.ME	0.908635	BANE.ME	-0.471528
18	TRNFP.ME	MSNG.ME	0.664100	AKRN.ME	-0.458741

Таблица 3 (сравнения корреляций акций)