Київський національний університет імені Тараса Шевченка
Факультет комп'ютерних наук та кібернетики
Кафедра дослідження операцій

ВИПУСКНА КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА
на здобуття ступеня бакалавра
за спеціальністю 113 «Прикладна математика»
на тему:

Задача сегментації ринку методами статистичного аналізу

Грабар Валерії Олексіївни
студентки 4 курсу

Науковий керівник:
доцент, кандидат фізико–математичних наук
Лівінська Г. В.

Робота заслухана на засіданні кафедри дослідження операцій та рекомендована до захисту в ЕК, протокол № 9 від 23 травня 2023 р.

Завідувач кафедри ДО  проф. Іксанов О.М.

Київ – 2023

РЕФЕРАТ

Обсяг роботи 64 сторінки, 25 ілюстрацій, 3 таблиці, 10 використаних джерел, 1 додаток.

ЗАДАЧА СЕГМЕНТАЦІЇ РИНКУ, КЛАСТЕРНИЙ АНАЛІЗ, МЕТОД K-MEANS, МЕТОД ВАРДА, МЕТОД DBSCAN, ПОРІВНЯННЯ МЕТОДІВ КЛАСТЕРИЗАЦІЇ.

Об’єктом дослідження є методи кластеризації та їх реалізація і порівняння задля вирішення маркетингової задачі сегментації ринку споживачів.

Метою роботи є теоретичне висвітлення різних методів кластеризації, а також їх практична реалізація та аналіз, задля знаходження найкращого сегментування ринку споживачів.

Інструменти розробки: середовище програмування PyCharm 2022.2, мова програмування Python 3.10.

У роботі виконаний теоретичний огляд етапів та методів задачі сегментації ринку, детально розглянуто один з її методів – кластерний аналіз, його методи та їх переваги і недоліки, розроблено програмне забезпечення для реалізації декількох методів кластеризації та попередньої обробки даних, зроблено висновки щодо результатів та отриманих кластерів.

ЗМІСТ

ВСТУП

Маркетингова стратегія – це комплексний план дій компанії, що спрямований на досягнення поставлених цілей в області маркетингу, зокрема, на збільшення прибутку та просування товарів чи послуг.

В умовах жорсткої конкуренції на ринку сьогодення, широке розуміння потреб споживачів та рівня їхньої задоволеності продуктом чи послугою є ключовими складовими успішної маркетингової стратегії. Для більшої ефективності у взаємодії зі споживачами, необхідно правильно розробити чи відрегулювати стратегію компанії.

Головними задачами маркетингу вважаються:

- Визначення цільових ринків
- Ринкове позиціювання
- Сегментування ринку
- Планування маркетингової діяльності

Задача сегментації ринку – одне з найважливіших завдань для компанії, воно полягає у поділі на групи схожих за своїми потребами та вподобаннями споживачів для подальшого налаштування маркетингової стратегії. Сегментація ринку є однією з ключових задач у маркетинговій діяльності, вона дозволяє компаніям відрізнити свій продукт від продукту конкурентів та ефективно взаємодіяти зі споживачами, пропонуючи їм товари та послуги, що підібрані за індивідуальними параметрами та потребами. Якщо коротко, метою сегментації ринку є вибір одного чи декількох сегментів споживачів, на задоволення потреб яких буде направлена стратегія компанії.

Для правильної сегментації ринку, компанії обирають для себе найбільш підходящий метод, залежно від їх потреб. Найпоширенішими є демографічний, географічний, психологічний та поведінковий методи. Кожен з них базується на різних критеріях вибору групи споживачів, що в свою чергу дає оптимальну стратегію для компанії.

Застосування статистичних методів для сегментації ринку дає змогу компаніям проводити це дослідження на основі великої кількості даних та використовувати їх для формування маркетингової стратегії. Хоча існує багато методів сегментації ринку, застосування статистичних методів вимагає відповідних знань та навичок з аналізу даних та моделювання. Для вирішення цієї задачі можна застосовувати різні статистичні методи, такі як факторний аналіз, кластерний аналіз, дискримінантний аналіз та інші, залежно від конкретної компанії та її цілі.

У даній кваліфікаційній роботі буде розглянуто застосування кластерного аналізу задля вирішення задачі сегментації ринку, досліджено теоретично та практично різні методи, які базуються на різних підходах, а також порівняння та аналіз результатів та висновки щодо отриманих сегментів ринку.

Доповідь структурована наступним чином. Розділ 1 – теоретичний, він включає в себе огляд усіх необхідних понять, розгляд маркетингових стратегій, поетапний розгляд задачі сегментації ринку, основні методи її вирішення, а також детальний огляд методів кластерного аналізу, алгоритми їх роботи, переваги та недоліки. Розділ 2 – практичний. Він включає в себе огляд задачі сегментації ринку конкретного випадку, огляд та попередню обробку даних, реалізацію різних методів кластеризації, порівняння їх результатів та висновок щодо отриманого результату.

У висновках наведено основні результати, отримані протягом виконання задачі сегментації ринку та порівняння методів кластеризації.

Мета роботи:

Дослідити теоретично різні методи вирішення задачі сегментації ринку, а також практично перевірити їх якість на конкретних даних.

Завдання роботи:

1. Повести теоретичний огляд задачі сегментації ринку
2. Розглянути етапи проведення задачі
3. Розглянути різні методи задачі сегментації ринку

4. Детально розглянути різні алгоритми кластеризації для вирішення задачі сегментації ринку
5. Реалізувати деякі методи кластеризації на конкретних даних
6. Порівняти їхні результати, обрати найкращий

РОЗДІЛ 1. ТЕОРЕТИЧНА ЧАСТИНА

1 РОЗГЛЯД МАРКЕТИНГОВИХ СТРАТЕГІЙ

Маркетингова стратегія – це довгостроковий план дій компанії для досягнення своїх маркетингових цілей. Він включає в себе визначення продукту чи послуги компанії, її місце на ринку, ціноутворення, її цільової аудиторії та просування.

В першу чергу, маркетингова стратегія визначається за типом призначення. Виділяють такі напрями:

1. Стратегія орієнтована на ціну: ця стратегія базується на конкурентоспроможній ціновій політиці. Компанії, які використовують цю стратегію, пропонують низьку ціну на свої продукти, що дозволяє їм конкурувати з іншими компаніями на ринку.
2. Стратегія орієнтована на продукт: суттю цієї стратегії є розробка високоякісних продуктів та послуг, які задовольняють споживачів. Здебільшого цю стратегію використовують ті компанії, які мають ресурси на дослідження та розробку продуктів.
3. Стратегія орієнтована на споживача: ця стратегія побудована на аналізі поведінки та потреб споживача. Компанії, які використовують цю стратегію, розробляють свої продукти з урахуванням бажань та потреб своїх клієнтів.
4. Стратегія орієнтована на конкурентів: це підхід, коли компанія спрямовує свої зусилля на аналіз та вивчення своїх конкурентів, та створенні продукту, кращого ніж у них.

Далі, залежно від головного орієнтира компанії, чи то споживач, чи то продукт, компанія шукає для себе оптимальну стратегію. На практиці найпопулярнішим напрямком стратегії є продукт. Передові компанії по всьому світу докладають чимало зусиль для підвищення якості продукту, збільшення

асортименту та підтримки інновацій. Наприклад, компанія Apple використовує таку стратегію, щоб створювати ідеальні продукти, які мають високий рівень якості та здатні задовольняти потреби споживачів. Прикладом стратегії орієнтовану на ціну є Walmart, що продає свої товари та послуги за низькою ціною порівняно з конкурентами.

Якщо ж розглядати конкретні маркетингові стратегії, які активно застосовуються на ринку, то найбільш популярними є:

1. Стратегія росту (Growth Strategy) – має на меті збільшення обсягу продажів та розширення географічного покриття.
2. Стратегія лідера вартості (Cost Leadership Strategy) – полягає в тому, щоб зайняти найнижчу цінову позицію на ринку і пропонувати товари чи послуги за нижчою ціною, ніж конкуренти. Це можливо завдяки ефективному управлінню витратами.
3. Стратегія диференціації (Differentiation Strategy) – передбачає створення унікального продукту чи послуги, які мають характеристики, що відрізняють їх від конкурентів. Компанії зосереджуються на високій якості продукті, унікальному дизайні чи підвищеному рівні обслуговування.
4. Стратегія фокусування (Focus Strategy) – полягає в зосередженні уваги на певному сегменті ринку з невеликою конкуренцією, або на певному регіоні. Ця стратегія дозволяє компанії зосередитись на конкретній групі споживачів та задовольнити їх потреби.

Ці стратегії на практиці довели свою ефективність, тому багато світових лідерів використовують самі їх для зростання та підтримки своїх бізнесів. Стратегія диференціації, наприклад, застосовується в Apple, Starbucks, Mercedes-Benz тощо. Такі компанії як IKEA, Walmart чи Ryanair зупинились на стратегії лідера вартості, бо продають продукти та послуги за нижчими цінами. Що ж до стратегії фокусування, її ми можемо знайти у таких гігантах як Ferrari (спорткари), Red Bull (енергетичні напої) чи Rolls-Royce (авіаційна двигуни).

Кожна з цих багатомільйонних компаній вже знайшла свою ідеальну маркетингову стратегію, і ми це бачимо через їхні прибутки, які щороку лише зростають. Але за кожним успішним проектом стоїть дуже багато важкої праці, досліджень, аналізу та спроб. Щоб знайти свій оптимальний шлях розвитку бізнесу, потрібно прикласти чимало зусиль, і одним з перших кроків, які потрібно проробити та дослідити є сегментація ринку.

2 СЕГМЕНТАЦІЯ РИНКУ. ЇЇ ЕТАПИ, ТИПИ ТА МЕТОДИ

Основною ціллю задачі сегментації ринку є поділ ринку на окремі групи споживачів, які мають схожі потреби, бажання, характеристики, поведінку чи інші фактори, які можуть впливати на їх вибір продукту чи послуги. Групи споживачів, що утворюються в результаті, називаються сегментами ринку. Кожен з сегментів має свої унікальні потреби та попит на товари та послуги.

Сегментація ринку допомагає компаніям більш ефективно взаємодіяти зі споживачами, забезпечуючи належну пристосованість продукту до їхніх потреб. Це дає змогу зменшити витрати на маркетинг, оскільки рекламні зусилля можуть бути спрямовані на конкретні сегменти ринку, які мають найбільший потенціал для покупок. Крім того, сегментація ринку допомагає збільшити задоволеність споживачів від продукту, що може позитивно відобразитися на репутації компанії та збільшити лояльність клієнтів.

Маркетологи стверджують, що ефективність сегментації ринку можна пояснити за допомогою принципу Паретто. Очевидно, що не всі споживачі ринку є рівними. Якщо слідувати поширеній версії закону Паретто, 20% споживачів роблять 80% прибутку, тоді як інші 80% споживачів роблять решту. Тому сегментування ринку якраз направлене на те, щоб знайти ці 20% аудиторії, зосередитись на них, скоротити витрати на маркетинг та отримати максимум виручки.

Виділяють п’ять основних етапів сегментації:

1. Визначення факторів сегментації
2. Вибір методу та здійснення сегментації
3. Розробка профілів груп споживачів
4. Оцінювання сегментів ринку
5. Вибір цільового ринку

2.1 Етап 1. Визначення факторів сегментації

Першим етапом є вибір ознак, за якими буде проводитись сегментація. Якщо говорити про сегментацію за групами споживачів, то існує декілька типів сегментації ринку, таких як демографічна, психографічна, географічна, поведінкова тощо. Кожен з них має свої особливості та методи визначення. Вибір методу сегментації залежить від типу продукту або послуги, що пропонується, та характеристик цільової аудиторії. Розглянемо детальніше найпоширеніші з них.

1. Географічна сегментація. Вона базується на розподілі споживачів за їхнім географічним розташуванням, таким як країни, регіони, міста тощо. Цей тип сегментації зазвичай використовують компанії, що продають товари або послуги, які мають залежність від місця проживання споживачів. Наприклад, підприємства, що продають товари побутової хімії, продукти харчування та інші товари, що вимагають фізичної присутності, можуть використовувати географічну сегментацію ринку.

Перевагами географічної сегментації ринку є змога компанії ефективно спрямовувати зусилля на певні регіони та місцевості, де попит на її продукцію є найвищим. Крім того, така сегментація може допомогти підприємству вивчати ринок та здійснювати ефективний маркетинговий аналіз.

Недоліками географічної сегментації ринку можуть бути високі витрати на рекламу та маркетинг у більш віддалених регіонах та нерівномірний розподіл споживачів, який може змінюватись в залежності від зовнішніх факторів.

2. Демографічна сегментація – це процес розділення ринку на групи покупців за допомогою демографічних характеристик, таких як вік, стать, рівень освіти, дохід, зайнятість, сімейний стан тощо. Цей тип сегментації є одним з найпоширеніших, оскільки демографічні характеристики є легкими для вимірювання та збору даних.

Переваги демографічної сегментації полягають у тому, що це дозволяє компанії зорієнтуватися на конкретні групи споживачів та розробляти маркетингові стратегії, які відповідають їхнім потребам. Крім того, демографічна

сегментація дозволяє підібрати підходящий медіа-мікс, що відповідає цільовій аудиторії. Наприклад, якщо цільова аудиторія старше 50 років, то медіа-мікс може бути націлений на телевізійні рекламні ролики та пресу, оскільки ці люди складають більшу частину аудиторії цих медіа.

Також однією з переваг цього методу є відкидання нецільової аудиторії, або зосередження на тих споживачах, які найбільш імовірно принесуть компанії прибуток. Наприклад, фокус на групі людей, які мають дохід вище певної суми грошей.

Недоліки демографічної сегментації полягають у тому, що вона не завжди дає повне уявлення про споживачів та їхні потреби. Наприклад, дві людини однакового віку та статі можуть мати різні інтереси та поведінку при покупках. Крім того, деякі демографічні характеристики можуть бути нестійкими, наприклад, дохід, який може змінюватися протягом часу.

3. Поведінкова сегментація враховує такі характеристики споживачів, як їх звички, в тому числі споживчі, мотивації та сприйняття продукту. Ці показники визначають, які продукти будуть використовуватись, які рекламні повідомлення будуть ефективними та який маркетинговий підхід буде найбільш успішним.

Переваги поведінкової сегментації ринку полягають в тому, що вона дозволяє компанії зосередитись на конкретній поведінці споживачів, а не на їх демографічних характеристиках. Це дозволяє створити більш ефективні маркетингові кампанії, оскільки вони будуть спрямовані на конкретну поведінку, що забезпечує більш точне позиціонування продукту та збільшує шанс його придбання споживачем.

Однак, поведінкова сегментація ринку має свої недоліки. Один з них полягає в тому, що вона вимагає від компанії більш складних досліджень і аналізу даних, оскільки необхідно визначити, які характеристики поведінки є найбільш важливими для визначення цільової аудиторії. Крім того, поведінкова сегментація ринку не враховує демографічні характеристики, які також можуть впливати на споживацьку поведінку.

Прикладом такої сегментації може бути виокремлення груп споживачів, за частотою їх покупок. Наприклад, компанія, яка продає товари для спорту, може визначити два різні сегменти: один з високою частотою покупок та високими ціновими преференціями, та інший навпаки - з низькою частотою покупок та низькими ціновими преференціями.

4. Психографічна сегментація: у цьому методі застосовуються дані про інтереси споживача, його цінності, стиль життя, переконання та поведінку. Цей метод дозволяє зрозуміти, що споживач очікує від продукту та як взаємодіє з ним.

Основною перевагою психографічної сегментації є те, що вона дає змогу пізнати більше про поведінку та потреби клієнтів, що в свою чергу може допомогти у розробці більш точної та ефективної маркетингової стратегії.

Недоліком психографічної сегментації може бути складність визначення психологічних характеристик цільової аудиторії, а також труднощі у визначенні, як саме ці характеристики пов'язані зі споживчою поведінкою.

Наприклад, компанія Apple використовує психографічну сегментацію, спрямовану на підлітків та молодих дорослих, які прагнуть бути модними та інноваційними. Їхня маркетингова стратегія спрямована на створення образу "крутого та незалежного" користувача, що сприяє збільшенню продажів для даного сегменту.

Також сюди можна додати приклади компаній, що виділяють сегменти споживачів, які піклуються про екологію, вегани чи притримуються правильного харчування. Зазвичай компанії при виборі правильної маркетингової стратегії обирають не одну, а декілька різних сегментацій. Більше прикладів поділів споживачів за певним критерієм наведемо нижче у таблиці 2.1:

Таблиця 2.1 - Типи сегментування ринку

Географічний	Демографічний	Поведінковий	Психологічний
Кліматичний аспект: - морський - континентальний - помірний -тропічний	Вік, стать: - 18 р., ч. - 25 р., ж	Статус споживача: - регулярний споживач - потенційний споживач	Стиль життя: - елітний - молодіжний - спортивний
Адміністративний: - країна - обласний центр - столиця - село	Сімейний статус: - одружений - незаміжня	Ступінь споживання: - слабкий - активний	Суспільний клас: - робітничий клас - середній клас
Населення: - більше 1 млн осіб - 500 тис – 1 млн осіб - 250 – 500 тис осіб - менше 5 тис осіб	Наявність дітей: - 2 дітей до 18р. - немає дітей	Очікувані вигоди: - ціна - сервіс - престиж	Тип особистості: - імпульсивний - конформіст
Регіони: - Закарпаття - Крим - Північ	Рівень доходу на місяць: - 5 тис грн - 30 тис грн - 250 тис грн	Ставлення до товару: - позитивне - байдуже - негативне	За адаптацією до нового товару: - консерватор - новатор
	Освіта: - вища, бакалавр - середня	Ступінь готовності придбати товар: - зацікавлений	

Якщо казати про сегментування промислового ринку, то використовуються такі групи факторів:

1. географічне розташування
2. розмір фірми
3. технології
4. галузь
5. очікуванні вигоди
6. інтенсивність споживання

Після того, як компанія визначилась з типом сегментація, вона переходить до наступного кроку – вибір методу.

2.2 Етап 2. Вибір методу для сегментування ринку

На даний момент існує два основних підходи до сегментування ринку. Перший підхід, що називається «a priory» (від лат. A priori - заздалегідь), має заздалегідь відомі такі ознаки сегментування: кількість сегментів, їх розмір, характеристики. Тобто в такому випадку сегментування є не основною задачею дослідження, а скоріше допоміжним засобом при виборі маркетингової стратегії. До прикладів апріорної сегментації можемо віднести такі поділи: чоловіки і жінки, дорослі і діти, південні та північні регіони.

Другий підхід, що має назву «post hos» (від лат. Post hoc - після цього), базується на те, що відсутня інформація щодо кількості сегментів та їх сутностей. У такому випадку сегменти визначаються після закінчення збору всієї необхідної інформації та аналізу усіх даних.

Основні методи, що використовуються в обох підходах, наведемо у таблиці 2.2:

Таблиці 2.2 - Підходи сегментування ринку

Підхід «a priory»	Підхід «post hos»		
	Напрям використання		
	Підготовка даних	Аналіз даних	Класифікація даних
Методи: - повний перепис верхнього шару споживачів - за обумовленими ознаками Критерії: Географічний, демографічний, психографічний, поведінковий	Методи: - факторний аналіз - конджойнт-аналіз - аналіз відповідності	Методи: - кластерний аналіз - нейронні мережі - CHAID аналіз - метод CART - латентний аналіз - часові ряди -дисперсійний аналіз	Методи: - дискримінантний аналіз - множинна регресія (Multiple Regression) -кластерний аналіз (якщо відома кількість кластерів)

Якщо говорити про класифікацію методів сегментування, можемо виділити однофакторні та багатофакторні. Загальну схему методів сегментації можна показати так (рис. 2.1)

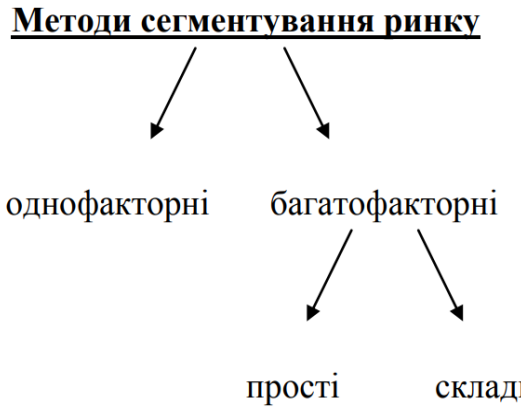

Рисунок 2.1 - Методи сегментування ринку

Суть однофакторних методів полягає у тому, що ми обираємо одну змінну, розбиваємо її на проміжки, інтервали чи значення, де відповідно кожне розбиття

відповідає окремому сегменту, і залежно від значення цієї змінної у споживача відносимо його до якогось сегменту. Такою цільовою змінною може бути стать, вік, зріст чи дохід споживача. Це вважається дуже простим підходом і не показує високої ефективності в більшості випадків, тому успішні компанії проводять сегментування ринку за одним з багатофакторних методів.

Багатофакторні методи мають в основі сегментування за декількома ознаками. Простим багатофакторним аналізом може бути такий, де поєднується два чи три демографічних показники. Що ж до складних методів багатофакторного аналізу, на практиці себе зарекомендували та проявили небагато методів, проте кожен з них доволі ефективний.

2.2.1 Метод угрупувань

Перший і найбільш поширений метод сегментації ринку за багатьма ознаками. Послідовно розбиваючи сукупності об’єктів на групи за їх найбільш вагомими ознаками, ми і отримаємо метод угрупувань. Тобто ми обираємо якусь ключову ознаку, далі за цією ознакою ми розбиваємо вибірку на дві частини (окремі підгрупи). Потім у кожній з цих підгруп ми знову обираємо нову найсуттєвішу ознаку (може бути різна для обох підгруп), за якою ми розіб’ємо ще на дві підрупи. Шляхом послідовного поділу на дві підгрупи, ми отримаємо низку підгруп із нашої початкової вибірки.

У іноземній літературі цей метод зустрічається під назвою метод AID, або ж Automatic Interaction Detection - метод автоматичного визначення взаємодій. Суть цього методу полягає у поступовій розбивці ринку на сегменти з послідовним дробленням їх залежно від заданого раніше набору критеріїв.

Наведемо приклад сегментування за цим методом на рисунку 2.2:

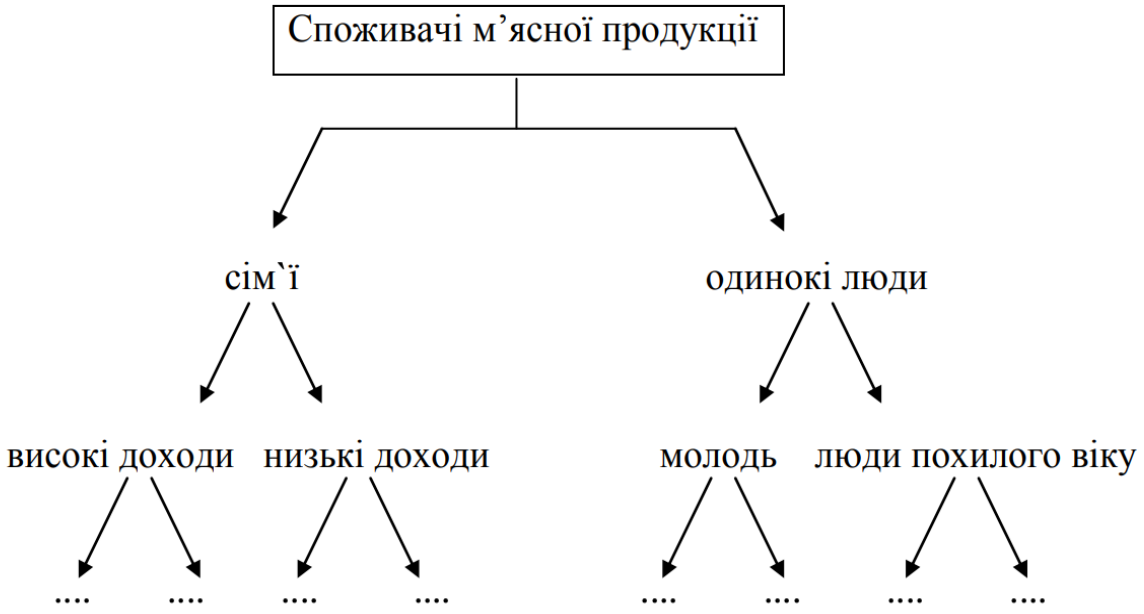

Рисунок 2.2 - Сегментування методом угрупувань

Відповідно до цього методу, першим етапом сегментування є аналіз усіх споживачів та вибір змінної, яка найбільше впливає на дохід компанії чи підприємства. У нашому прикладі виявилось, що найважливішою змінною є наявність у споживачів сім’ї, тому перший поділ на підгрупи відбувся саме за цією ознакою. Другим кроком є вибір найсуттєвішої змінної для тих споживачів, у яких є сім’я. У нашому випадку аналіз показав, що сім’ї з високим та низьким показниками доходу мають доволі різні характеристики, тому наступний поділ відбувся на базі цієї змінної. За цією ж логікою поділяємо одиноких людей за віковою групою. Повторюючи ці кроки для новостворених підгруп ми дійдемо до того моменту, коли у сегментах стане або замало споживачів, або стане неможливо виділити ключову ознаку.

Головною перевагою цього методу є його здатність швидко формувати сегменти, які суттєво відрізняються один від одного. До недоліків можна додати той факт, що сегментування цим методом не враховує можливість взаємодії ключових ознак між собою, бо ми розглядаємо одночасно лише один вимір сегментування.

2.2.2 Методи багатовимірного статистичного аналізу

Цей метод також має назву «метод багатомірної класифікації», його суть полягає в класифікації об’єктів за декількома ознаками одночасно. При застосуванні цього методу, найбільш ефективним вважається використання кластерного аналізу. Кластерний аналіз базується на двох принципах:

1. споживачі в одному кластері мають бути максимально схожими, а споживачі з окремих кластерів – максимально несхожими;
2. споживачі об’єднуються в один кластер, якщо мають кілька подібних ознак.

Цей спосіб використовується тоді, коли важко заздалегідь сказати, яка кількість кластерів має бути, чи яка кількість споживачів має належати до цих кластерів.

Чим відрізняється цей метод від методу угрупувань? Метод угрупувань має за основу принцип поступового поділу загальної вибірки на підгрупи, якщо ж говорити про кластерний аналіз, то тут спочатку розглядається кожен споживач окремо, а лише потім вони об’єднуються в підгрупи (кластери) за якимись схожими ознаками. Між собою також можуть об’єднуватись кластери, якщо вони мають схожі ознаки. Поділ на кластери припиняється тоді, коли більше немає схожих між собою об’єктів, які можна було б об’єднати.

У методі кластерного аналізу є певне обмеження, або радше рекомендація: щоб кластеризація була успішною, необхідно щоб у ній брало участь щонайменше 200 споживачів. Але трапляються випадки, коли і менша кількість об’єктів чудово кластеризується.

Головною перевагою методу кластерного аналізу є його простота та можливість охоплювати багато ознак одночасно. Недоліком можна назвати те, що при використанні цього методу існують ризики неправильного, чи неоднозначного формування кластерів коли не вистачає інформації або коли дані неправильно інтерпретуються. Саме тому при роботі із цим методом варто бути уважним,

приділити багато часу маркетинговим дослідженням та ретельно перевіряти отриманий результат, краще навіть на кожному новому кроці кластеризації.

2.2.3 Інші методи сегментації

Окрім описаних вище двох найпоширеніших методів сегментування ринку, існують також менш поширені.

Перший з цього переліку є метод побудови сітки сегментації. Цей метод використовується на рівні макросегментації, його основна функція – виділення базових ринків. При застосуванні методу побудови сітки ключовими вважаються три змінні:

- технології
- споживачі
- вигоди, які шукають споживачі (або функції)

У результаті отримують стратегічно важливі сегменти.

Цей метод використовують здебільшого ті підприємства, які ще не визначились, в якому регіоні та на якому ринку краще за все реалізовувати свою продукцію, або ті, які планують виходити на міжнародний ринок. В інших випадках використання цього методу не є доцільним.

До більш нових методів можна віднести метод функціональних карт. Цей метод проводить сегментацію за споживачем та за товаром, тому можна сказати, що цей метод заснований на «подвійній» сегментації. Якщо говорити про функціональні карти, то вони в свою чергу можуть бути такими:

- однофакторними (якщо сегментація відбувається за одним чинником та проводиться для однорідної групи товарів)

- багатофакторними (якщо аналізується, для якої групи споживачів призначена конкретна модель товару і які головні параметри цього товару для просування його на ринку); основними чинниками є ціна, технічні характеристики та канали збуту.

Початкові дані мають вигляд матриці, де рядки – значення чинників, а стовпці – сегменти. Принцип цієї матриці заключається в тому, що за її допомогою ми визначаємо, які параметри найбільше підходять для конкретних сегментів, тобто є ключовими метриками при сегментації.

Також до нових методів відноситься метод сегментації за вигодами, або гнучка сегментація. Цей метод, на відміну від інших, заснований на поведінці споживачів та врахуванні відмінностей в їх цінностях.

За основу в цьому методі взято поведінку споживачів, їх реакцію на певні споживчі ситуації, типу акції та знижки, різні реклами і тому подібне.

Споживачі розподіляються по сегментах залежно від категорії поведінки, демографічних показників, географії, способу життя та споживчих переваг.

Недоліком цього методу є те, що можливо лише помітити відмінність у реакціях споживачів на певні товари, але зрозуміти, чому була така чи інакша реакція неможливо.

Отже, можемо зробити невеликий висновок. На теперішній момент не можна обрати якийсь один універсальний метод сегментації ринку, який би підійшов усім та покрив усі вимоги, але як ми бачимо, кожен метод має свої переваги та недоліки, тому задача сегментації ринку для компанії починається з того, щоб обрати найоптимальніший для себе метод та отримати найбільш ефективні результати.

2.3 Етапи 3, 4, 5. Оцінювання та вибір результатів.

Після того, як підприємство визначилось із оптимальним методом сегментування ринку та здійснило його, потрібно правильно інтерпретувати отримані групи споживачів. Це і є третім етапом задачі сегментації ринку. На цьому кроці необхідно для кожного новоствореного сегменту виділити його головні ознаки, за якими відбувалось сегментування, та описати за цими ознаками цю групу споживачів.

Після опису сегментів, необхідно їх правильно оцінити. Це буде четвертим етапом сегментування ринку. Оцінювання сегментів теж непроста задача. Спочатку необхідно зрозуміти, наскільки гарно фірма зможе конкурувати у цьому сегменті та загалом вияснити, наскільки цей сегмент привабливий та перспективний.

Щоб правильно оцінити можливості фірми на конкуренцію у цьому сегменті, необхідно перевірити фірму на наявність певних характеристик:

- технологічні нововведення, яких не мають конкуренти
- достатні фінансові ресурси
- відповідна кваліфікація персоналу
- маркетингові можливості компанії
- адекватні вимоги ринку

Щоб зрозуміти, наскільки сегмент привабливий та вартий зусиль, необхідно звернути увагу на такі критерії:

- соціальні, політичні та екологічні фактори
- конкурентні фактори (наявність теперішніх та можливість нових конкурентів)
- ринкові фактори (темп зростання сегменту, його розмір, чутливість до цін та їх змін).

Загалом, опираючись на ці пункти, можна описати «ідеальний» сегмент для просування:

- швидкий темп зростання
- високий рівень прибутку
- адекватний рівень конкуренції

На жаль, на практиці така комбінація ознак майже неможлива. Адже маючи перші два фактори, сегмент навряд буде мати небагато конкуренції. Тому оцінка та вибір оптимального сегменту – це в першу чергу пошук компромісів.

Наступним та останнім етапом задачі сегментування ринку є власне вибір сегменту, на якому фірма зосередить свою увагу. Для того щоб його обрати, потрібно відштовхуватись від маркетингової стратегії.

Є два варіанти: диференційований маркетинг та недиференційований.

Стратегія другого обирається у випадку, якщо підприємство виходить на ринок з одним товаром, та концентрується на потребах споживачів, не зважаючи при цьому на їх відмінності. Тобто це може бути якийсь загальний товар, який потрібен широкому сегменту, куди входить багато категорій людей. Цей метод не є дуже популярним на сучасному ринку.

Диференційований маркетинг більш поширений. Його суть полягає у виборі компанією декількох сегментів та виготовленні для кожного з них окремих товарів та відповідно різних методів просування. Прикладом такої стратегії може бути виробництво молочної продукції, яке виробляє різні види молока: звичайне, знежирене, безлактозне, рослинне, що в свою чергу орієнтоване на різні сегменти споживачів. Стратегія диференційованого маркетингу теж може мати різні напрямки:

- повне охоплення ринку
- товарна сегментація (для різних сегментів пропонується однаковий товар)
- вибіркова спеціалізація (окремим сегментам – окремі товари)
- сегментна спеціалізація (одному сегменту пропонуються всі товари)

3 КЛАСТЕРНИЙ АНАЛІЗ

3.1 Основні поняття та принципи кластерного аналізу

Як уже згадувалось раніше, одним із найпоширеніших методів сегментації ринку є метод багатовимірного статистичного аналізу, зокрема метод кластерного аналізу. Окрім вже відомого, що кластерний аналіз – це поєднання багатовимірних статистичних методів, зазначимо, що це також статистична задача, суть якої полягає в об'єднанні схожих за характером елементів з конкретної вибірки, які разом утворюють самостійні одиниці, що мають певні властивості – кластери.

Взагалі кластерний аналіз з'явився відносно недавно – у 1939 році. Він був запропонований вченим К. Тріоном. Дослівно термін «кластер» перекладається з англійської як гроно, група, пучок. Свого найбільш бурхливого розвитку кластерний аналіз набув у 60-х роках минулого століття. Цьому сприяли поява швидкісних комп'ютерів, а також розуміння того, що класифікація – це фундаментальний метод наукових досліджень.

Метою кластерного аналізу є утворення з певної вибірки елементів декількох підгруп (кластерів), де всередині кожної з них будуть знаходитись максимально схожі між собою елементи, та одночасно з цим, елементи з різних кластерів мають максимально відрізнятися. Виходячи з цього, кластеризацію можна назвати задачею класифікації, адже в основі обох задач лежить поділ об'єктів на базі їх схожості, хіба у кластеризації наперед не задана приналежність певних об'єктів класам.

Враховуючи, що кластеризація є багатовимірним методом, можемо зробити висновок, що в її основі лежить поділ об'єктів за подібністю декількох ознак. Тобто розбивка на кластери відбувається не за однією метрикою, а за набором багатьох різних ознак. Для того щоб коректно виділити однорідні групи із загальної вибірки за декількома ознаками схожості, необхідно ввести певні показники, які характеризують ступінь близькості об'єктів за усіма параметрами.

Окрім самих параметрів, за якими визначається рівень подібності, грають роль ще алгоритми кластеризації. Вони бувають різні, залежно від того, за яким принципом відносити об’єкти до одного кластеру: за щільністю, за відстанню між ними або за розподілами. Це залежить від самих даних та цілі кластеризації.

Перед проведенням безпосередньо кластеризації необхідно вирішити декілька важливих речей:

- на базі яких показників чи метрик буде проводитись кластеризація. Це питання є важливим, оскільки від цього залежить правильність розбиття на кластери та що ці кластери будуть характеризувати.

- Яким методом проводити кластеризацію. Для цього треба гарно розумітися на усіх (або хоча б основних) методах кластеризації, знати принципи їх роботи, переваги та недоліки, щоб обрати саме той метод, який буде підходити найбільше.

- На скільки кластерів розбивати вибірку. Якщо така інформація невідома заздалегідь, необхідно провести ряд досліджень аби знайти оптимальну кількість груп.

- Як правильно інтерпретувати результати. Потрібно переконатись, що кластеризація проведена правильно, що не утворилось зайвих кластерів або кластерів, у яких насправді між об’єктами немає нічого спільного, а об’єднує їх лише велика відмінність від усіх інших кластерів. Також важливо впевнитись, що утворені кластери дійсно сформувались як і потрібно, адже інколи у початковій вибірці насправді немає ніяких кластерів.

Формальна постановка задачі має такий вигляд:

Нехай у нас є вибірка об’єктів $X^t = \{x_1, \dots, x_t\} \subset X$ та певна функція відстаней між цими об’єктами $\rho(x, x)$. Задача полягає у розбитті вибірки на підмножини, що не перетинаються та складаються з об’єктів, між якими мала відстань по метриці ρ. Також кожному об’єкту $x_i \in X^t$ ставиться у відповідність y_i - номер кластеру.

Функція $a: X \to Y$ - це алгоритм кластеризації, який призначає кожному об’єкту $x \in X$ мітку кластера $y \in Y$. Інколи число елементів множини Y відомо

наперед, але здебільшого це є ще однієї задачею – визначити оптимальну кількість кластерів.

Подібність об’єктів, за якою вони і потрапляють до одного кластеру – це фактично кількісне значення, яке показує ступінь взаємозв’язку між об’єктами.

Взагалі кластеризація – це доволі поширена задача, яка використовується в багатьох сучасних сферах. Наприклад, пошук інформації, машинне навчання, комп’ютерна графіка чи розпізнавання образів, але в основному її застосовують для групування даних.

За своїм типом, кластеризація поділяється на м’яку та жорстку. Остання означає, що об’єкт або належить якомусь кластеру повністю, або не належить зовсім. У першому типі кластеризації – м’якій – об’єкт може належати до кластера з певною імовірністю. Тоді ж об’єкт може належати до декількох кластерів, сума імовірностей яких дорівнює одиниці.

Якщо говорити про типи даних, які приходять на вхід, часто вони не є одного виду. Це може бути суміш категоріальних, інтервальних та порядкових даних. У такому випадку кластеризація стає більш важким процесом, адже від типу даних залежить спосіб вимірювання відстані між об’єктами. Тоді щоб все-таки правильно виміряти відстань між досліджуваними об’єктами, наприклад для інтервальних даних використовують Евклідову відстань.

Якщо ж у нас всі дані мають один вид, вони можуть бути у різних шкалах. Тоді потрібно провести нормування даних, щоб усі метрики мали однакову вагу при кластеризації. Перелічимо основні способи нормування даних:

1. Максимум 1: значення змінних діляться на їх максимум.
2. Середнє 1: значення змінних діляться на їх середнє.
3. Z-Scores: від кожного значення змінної віднімається їх середнє та ділиться на стандартне відхилення.
4. Розкид від 0 до 1: за допомогою лінійних перетворень усі змінні стають в діапазоні від 0 до 1.
5. Розкид від -1 до 1: за допомогою лінійних перетворень усі змінні стають в діапазоні від -1 до 1.

Що ж до самих кластерів, можемо виокремити декілька різних типів:

1. Відокремлені кластери. Кожен елемент кластера має менші відстані до елементів у своєму кластері, аніж до елементів у інших кластерах.
2. Кластери на базі центрів. Елемент знаходиться ближче до центру свого кластеру, аніж до центрів інших кластерів.
3. Кластери з сусідством. Кожен елемент має меншу відстань як мінімум до одного елементу свого кластеру, ніж до будь-якого іншого елементу інших кластерів.
4. Кластери на основі щільності. Тут кластери – це ділянки з високою щільною, відділені від інших кластерів ділянками із низькою щільністю.
5. Концептуальні кластери. Елементи у кластері мають спільну властивість, яка походить від усієї вибірки елементів.

Якщо говорити про математичні характеристики кластерів, можна назвати такі:

1. Центр – середнє геометричне місце точок у просторі змінних.
2. Радіус – максимальна відстань від центру кластера до будь-якої його точки.
3. Діаметр – максимальна відстань між двома будь-якими точками кластера.
4. Щільність – відношення кількості точок кластера до радіуса в якійсь степені.
5. Розмір кластера – кількість точок у кластері.
6. Середньоквадратичне відхилення.
7. Міжкластерна відстань – відстань між центрами кластерів, між найближчими точками різних кластерів, середня відстань між усіма парами точок з різних кластерів.

Серед математичних характеристик було згадано центр кластера. Центр кластера може бути двох видів, залежно від простору:

- якщо простір Евклідовий, то центром кластера, або центроїдом, вважається середнє арифметичне координат точок кластера

- у неевклідовому просторі центроїда немає. Натомість центром кластера обирається одна з точок кластера, що мінімізує суму відстаней до центра, суму квадратів відстаней та максимальну відстань до інших точок з інших кластерів.

Також не менш важливим є питання зупинки кластеризації. Для правильного моменту існують кілька критеріїв:

- за час останньої ітерації кластери не змінились

- поділ дійшов до потрібної кількості кластерів

- математичні характеристики кластерів досягли граничних значень.

Отже, коли вже відомі загальні теоретичні відомості, логічним наступним кроком є саме вибір правильного методу кластеризації.

3.2 Методи кластеризації

На даний момент існує багато різних алгоритмів для кластеризації. Кожен з них має свій принцип роботи, розглянемо основні з них, які найчастіше використовуються на практиці.

3.2.1 Ієрархічні методи кластеризації

Першою класифікацією методів буде поділ на ієрархічні та неієрархічні. Перші об’єднують об’єкти у ієрархічні фази, звідси власне і походить назва. Відповідно одночасно змінює свою групу лише один об’єкт, а інші залишаються у своїх кластерах. Зазвичай результат ієрархічних методів представляється у вигляді дендрограми.

Дендрограма – це тип деревної діаграми, яка показує ієрархічні взаємозв’язки між різними наборами даних. (Рис. 3.1)

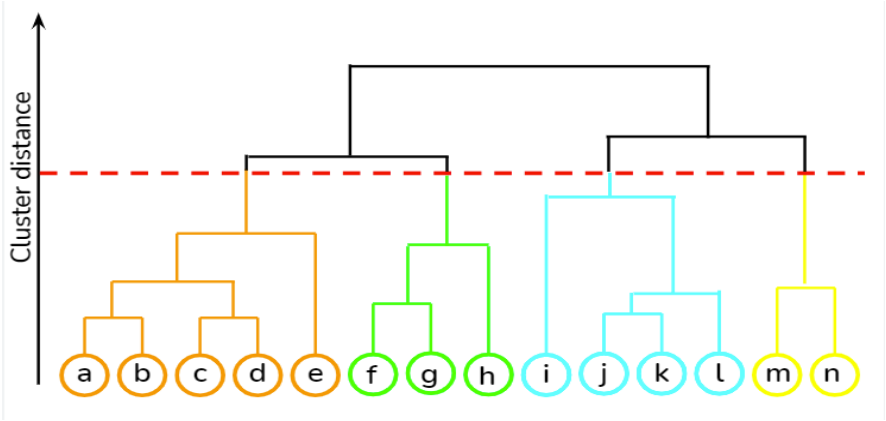

Рисунок 3.1 - Дендрограма

Дендрограма представляє собою декілька важливих речей:

- відстань між точками означає їх несхожість

- висота блоків означає відстань між кластерами

- по малюнку можна зрозуміти, як формувався кластер.

Також важливою задачею дендрограми – є зображення кластерної структури у вигляді графіка у двох вимірах, незалежно від того, якою була початкова розмірність простору. Звісно, є й інші способи візуалізувати багатовимірні дані. Наприклад, карти Кохонена, але такі методи зазвичай привносять у картину штучні спотворення.

Основною задачею ієрархічних методів є побудова ієрархії кластерів. Цей процес відбувається автоматично: згори-вниз або знизу-вгору, залежно від алгоритму. В будь-якому випадку, перед початком роботи алгоритму необхідно задати бажану кількість кластерів та умову завершення.

В свою чергу, ієрархічні методи кластеризації поділяються на агломеративні та дивізійні. Принцип агломеративних (Agglomerative Nesting, AGNES) методів заключається у тому, що на початку роботи алгоритму кожна точка вибірки вважається окремим кластером, які на кожному кроці об’єднуються в більші кластери. Це об’єднання відбувається на базі відстані між точками чи кластерами. Кінцевою метою цього методу є об’єднання усіх точок в один великий кластер.

Найвідомішими представниками таких методів є:

1. Метод Варда: цей метод є найпопулярнішим серед ієрархічних методів ще з 1963 року. Він використовується для послідовного об'єднання найближчих кластерів на кожному кроці з метою мінімізації зростання внутрішньокластерної дисперсії. У результаті утворюються кластери приблизно одного розміру, що мають форму гіперсфер.
2. Метод найближчого сусіда: у цьому методі робота алгоритму заключається у пошуку мінімальної відстані між найближчими точками. Цей метод доволі чутливий до даних, що можуть спричинити «шум». Подібний до методу дальнього сусіда.
3. Метод середніх: цей метод використовується для послідовного об’єднання найближчих кластерів на кожному кроці на основі середньої відстані між кластерами. Цей метод допомагає зменшити «шум».

Наступною групою ієрархічних методів є дивізійні (Divisive Analysis, DIANA) методи, суть яких протилежна агломеративним – на початку роботи у нас є один кластер, який на кожному кроці поділяється на менші і вкінці-кінців розбивається на кожну окрему точку. Кожен з агломеративних методів застосовний також до дивізійного принципу кластеризації. Схематичне зображення роботи агломеративного та дивізійного принципів зображено на рисунку 3.2

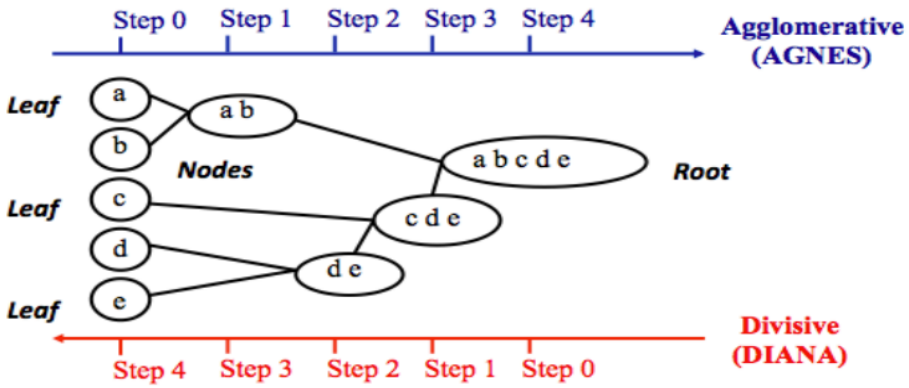

Рисунок 3.2 - Ієрархічні методи кластеризації

3.2.2 Неієрархічні методи кластеризації

На відміну від вище описаних ієрархічних методів, неієрархічні є більш стійкими до шумів та викидів, до неправильного вибору метрики та до зайвих змінних, що беруть участь в кластеризації. Проте, тут є і своя ціна. Аналітик має знати кількість кластерів або умову зупинки ще до початку роботи алгоритму.

З-поміж неієрархічних методів важливо виділити ітеративні методи. Логіка їх алгоритмів заключається в наступному:

1) Початкова вибірка даних розбивається на якусь кількість кластерів. Після цього для кожного кластеру обчислюється його центр тяжіння.
2) Після цього кожна точка з вибірки потрапляє в той кластер, чий центр тяжіння є найближчим до неї.
3) Наступним кроком є переобчислення центрів тяжіння для кожного кластера.
4) Повторення кроків 2 і 3, доки кластери не перестануть змінюватись.

Загальний принцип роботі ітеративних методів зображено на рисунку 3.3

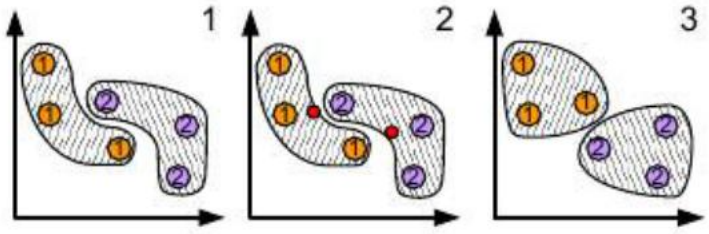

Рисунок 3.3 - Схема роботи ітеративних методів

Враховуючи той факт, що для ієрархічних методів необхідно обчислювати матрицю схожості між об’єктами розмірністю $N \times N$, а ітеративні методи цього не потребують, за допомогою останніх можна обробляти значно більші обсяги даних.

Першим та найпопулярнішим представником неієрархічної кластеризації є метод k-середніх (K-Means), де k означає кількість кластерів. Цей метод майже одночасно винайшли у 1950-х роках Гуґо Штайнгауз та Стюард Ллойд, проте найбільшої популярності він отримав після того, як Джеймс МакКвін випустив

свою роботу «Some methods for classification and analysis of multivariate observations», де вперше вжив термін «k-середні».

Метою методу k-середніх є розподілення n об'єктів по k кластерах, так щоб кожен об'єкт належав до кластеру з найближчим до нього середнім значенням. По суті, цей метод є чисельною реалізацією методу найменших квадратів, адже основною задачею цього алгоритму є мінімізація середньої квадратичної відстані між точками в одному кластері, тобто функції

$$\sum_{i=1}^{n} d(x_i, m_j(x_i))^2$$

де d – Евклідова відстань, x_i – i-тий об'єкт у вибірці, m_j- центр кластера, якому був призначений об'єкт x_i на j ітерації.

Центроїд кластеру можна описати за допомогою наступної формули:

$$C_i = \frac{1}{\|S_i\|} \sum_{x_j \in S_i} x_j$$

де C_i – i-тий центроїд, S_i – всі точки, що належать кластеру i, x_j – j-та точка з вибірки, $\|S_i\|$ – кількість точок в кластері i.

Сам алгоритм має такий вигляд:

1. Визначаємо, скільки у нас буде створено кластерів – k.
2. За допомогою рандому обираємо k точок із початкової вибірки, що стануть початковими центроїдами.
3. Кожну точку вибірки, окрім центроїдів, віднести до якогось кластера, де відстань від точки до центроїда буде найменшою Евклідовою відстанню.
4. Для кожного кластеру перераховуємо його центроїд за допомогою середніх значень усіх точок у кластері.
5. Повторюємо кроки 3 і 4, доки наші кластери не перестануть змінюватись, або алгоритм досягне максимальної кількості ітерацій.

Метод має багато переваг, таких як швидкість та зручність для великої кількості даних, проте він є чутливим до викидів, будує кластери лише опуклої форми та працює лише з числовими даними. Не дивлячись на його широке

застосування, все ж є випадки, коли цей метод не справляється із правильною кластеризацією. Такі випадки представлені на рисунку 3.4

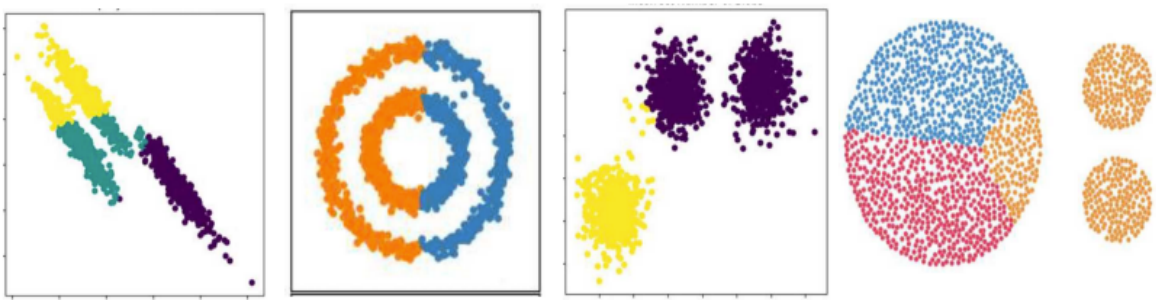

Рисунок 3.4 - Неправильна кластеризація методом k-середніх

Також однією з основних проблем методу k-середніх є початкова ініціалізація центроїдів. Як вже згадувалось, вони обираються випадковим чином із усієї вибірки елементів, проте часто початкове задання центроїдів впливає на результат кластеризації. На рис. 3.2.5 можна побачити два різних варіанти початкових центрів та власне різні кластери

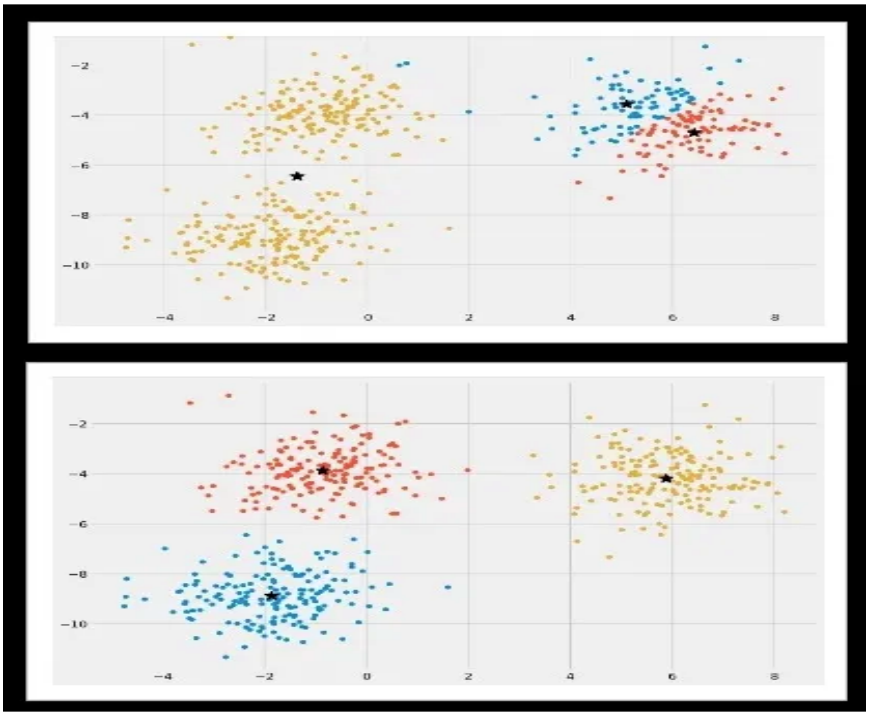

Рисунок 3.5 - Початкові центроїди та результати кластеризацій

Щоб вирішити цю проблему, можна використати метод k-means++. K-Means++ - це розумна техніка ініціалізації центроїдів, а решта алгоритму нічим не відрізняється від методу K-Means.

Вибір початкових центроїдів має такий алгоритм:

1. Обираємо перший центроїд C_1 випадковим чином.

2. Обчислюємо відстань від усіх точок вибірки до даного центроїда. Максимальну відстань точки x_i від центроїдів можна зобразити за допомогою формули

$$d_i = max_{(j:1\rightarrow m)} \left\| x_i - C_j \right\|^2$$

де d_i – відстань точки x_i від найдальшого центроїда, m – кількість уже обраних центроїдів

3. Робимо точку x_i новим центроїдом
4. Повторюмо кроки 2 і 3 доки не буде обрано k центроїдів.

Вже маючи зрозумілий метод першої ініціалізації центроїдів, залишається розібратись із один – скільки ж має бути кластерів. Іноді ці інформація вже відома аналітику відштовхуючись від якихось логічних міркувань, якщо він гарно знайомий з даними чи продуктом. Але ж трапляються ситуації, коли визначити кількість кластерів – це окрема повноцінна задача, для вирішення якої існують окремі методи.

Один з таких методів – метод Ліктя (Elbow Method). Цей метод має на увазі під собою багаторазове виконання кластеризації з подальшими збільшенням кількості кластерів та відкладанням на графіку оцінки кластеризації, обчисленої як функція від кількості кластерів. Якщо точніше, на кожній ітерації ми обчислюємо дисперсію всередині кластерів. Відповідно, коли у нас буде всього один кластер, у нас буде найбільша дисперсія, а коли у нас буде стільки кластерів, скільки й елементів у вибірці, дисперсія буде мінімальною. Наша задача на графіку із цих дисперсій знайти момент, коли їх значення починає спадати дуже повільно, а не так стрімко як на початку.

На рисунку 3.6 зображено приклад графіку, де по вертикалі у нього внутрішньокластерні дисперсії, а по горизонталі – кількість кластерів.

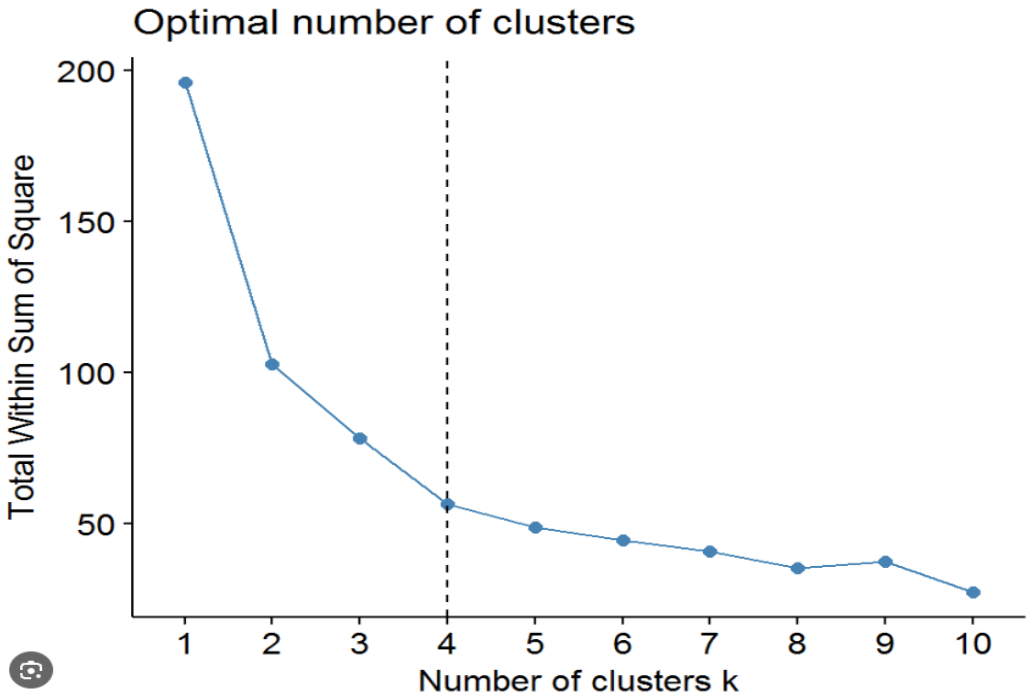

Рисунок 3.6 - Метод ліктя

Загалом, можна зрозуміти, що тут оптимальна кількість кластерів – 4, але це не зовсім очевидно, адже перепад між 4 і 5 та 5 і 6 майже однаковий. Щоб бути більше впевненим, можна використовувати Метод Силуетів.

Метод Силуетів був запропонований бельгійським статистиком Пітером Руссеу в 1987 році. Цей метод демонструє, наскільки близько кожна точка в одному кластері знаходиться до точок в сусідніх кластерах. Коефіцієнт «силуету» має значення від -1 до 1 та обчислюється за допомогою внутрішньокластерної відстані (a) та середньої відстані до найближчого кластера (b) по кожному об'єкту:

$$\frac{b-a}{\max(a,b)}$$

Графік Методу Силуету зображує кластери, але у незвичній для нас формі. Чим вищий кластер, тим більше елементів у ньому, чим ближче він знаходиться до 1, тим більше він ізольований та віддалений від інших кластерів. На рисунках 3.7 та 3.8 зображені графіки Методу Силуету для 2 та 4 кластерів, де червона лінія – середнє значення коефіцієнту Силуету.

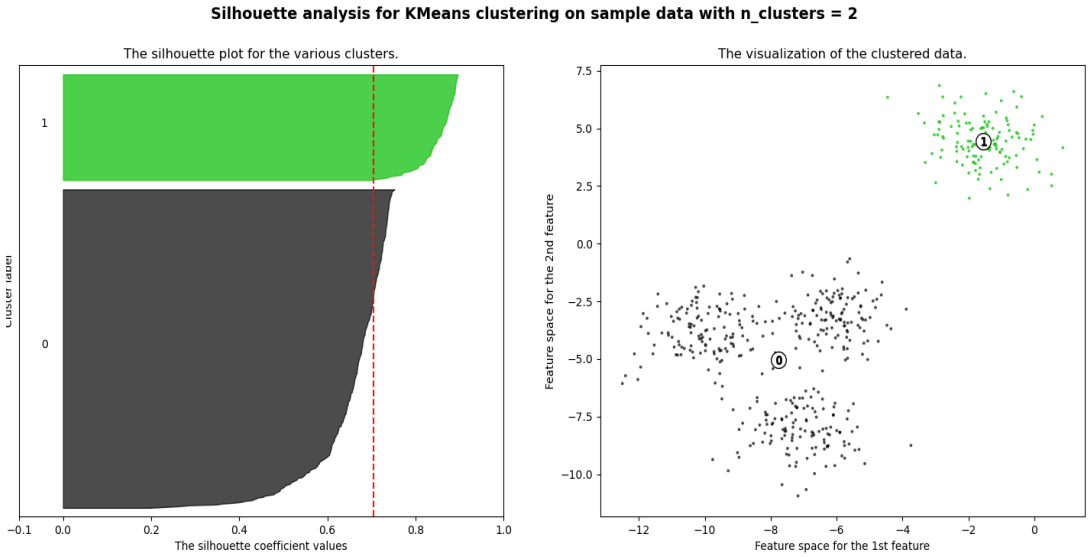

Рисунок 3.7 - Метод Силуету для 2 кластерів

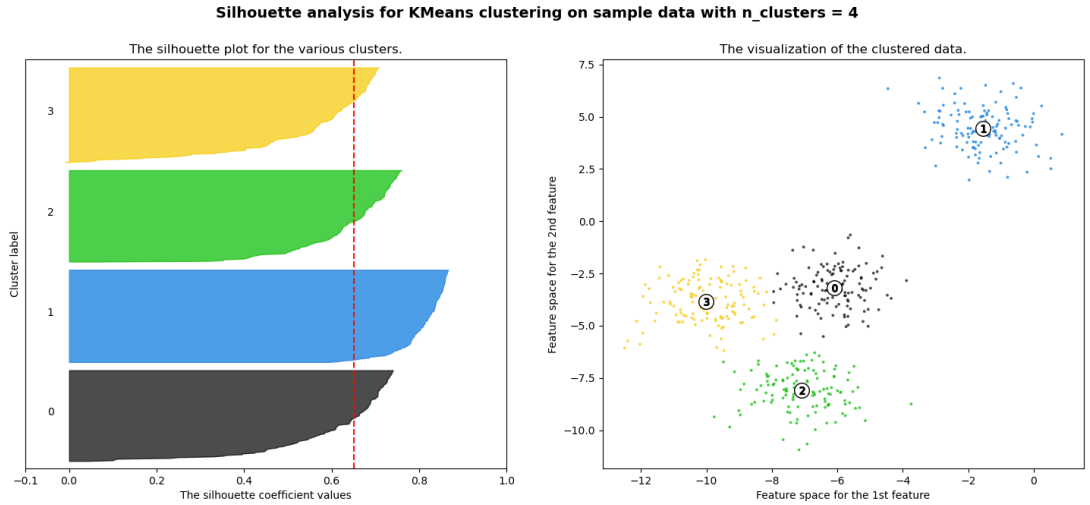

Рисунок 3.8 - Метод Силуету для 4 кластерів

Щоб з'ясувати, яка кількість кластерів оптимальна за Методом Силуету, потрібно ітеративно запустити його для різної кількості кластерів і на кожній ітерації обчислювати середнє значення коефіцієнту Силуету. Для оптимальної кількості кластерів це значення буде найбільшим.

Проте, навіть правильно заздалегідь визначився кількість кластерів та правильно ініціалізувавши центроїди, не вийде уникнути усіх проблем із методом k-середніх. Наприклад, щоб правильно визначити «вкладені» кластери, можна використовувати методи, що засновані на щільності між точками. Яскравим прикладом таких методів є DBSCAN. Його назва розшифровується як Density Based Spatial Clustering of Application with Noise, що вказує на кластеризацію за

щільністю з шумом. Цей алгоритм запропоновали Ганс-Петер Крігель, Сяовей Су, Йорг Сандер та Мартін Естер у 1996 році. Цей метод, на відміну від ієрархічних методів та K-Means, дуже ефективно себе показує із кластерами нестандартної форми, а також виділяє шуми та викиди в окремі кластери.

Основна ідея DBSCAN полягає в тому, що точка належить кластеру, якщо вона близька до багатьох точок цього кластеру. У цьому алгоритмі присутні два головних параметри:

1. eps – відстань, що визначає радіус. Дві точки вважаються сусідніми, якщо відстань між ними менше або дорівнює eps.
2. minPts – мінімальна кількість точок для визначення кластера.

На основі цих двох параметрів, точки у вибірці класифікуються як основні точки, граничні точки або викиди:

- Основна точка: це така точка, в оточенні якої з радіусом eps є принаймні minPts точок (включаючи саму точку).

- Гранична точка: точка є граничною, якщо основна точкою є сусідньою до неї, але в околі точки менше ніж minPts точок.

- Викид: точка вважається викидом, якщо вона не є основною, а також жодна точка не є сусідньою для неї.

Ці види точок зображено на рисунку 3.9, де A – основна точка, B – гранична точка, N – викид, minPts = 4

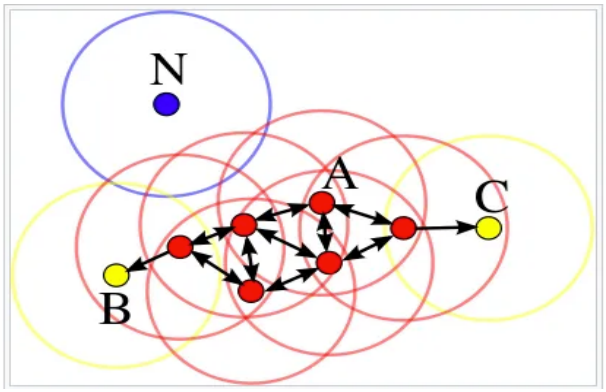

Рисунок 3.9 - Види точок в алгоритмі DBSCAN

Тепер більше про сам алгоритм роботи методу DBSCAN:

- визначаються minPts та eps.

- Початкова точка обирається випадковим чином, її окіл визначають за допомогою радіуса eps. Якщо в околі точки є принаймні minPts точок, ця точка позначається як основна і починається формування кластера. Якщо ні, точка позначається як шум. Після початку формування кластера (нехай кластер А), усі точки в околі початкової точки стають частиною кластера А. Якщо ці нові точки також є основними, точки, які знаходяться поблизу них, також додаються до кластера А.

- Наступним кроком є випадковий вибір іншої точки серед тих точок, що не були відвідані на попередньому кроці. Потім застосовується та сама процедура.

- Цей процес завершується, коли всі точки відвідані.

Варто зазначити, що відстань між точками вимірюється за допомогою Евклідової відстані, як і в методі k-середніх. Застосовуючи ці кроки, алгоритм DBSCAN може знайти області високої щільності та відокремити їх областями з низькою щільністю.

У цьому методі кластер включає основні точки, що є сусідніми, та всі їх граничні точки. Необхідною умовою для формування кластера є наявність принаймні однієї основної точки. І хоча це малоймовірно, може утворитись кластер, що складатиметься із однієї основної точки та її граничних точок. Повертаючись до рис. 3.2.4 із випадками кластеризації, коли метод k-середніх не справився б з ними, метод DBSCAN зміг би правильно виконати кластеризацію на кожному з цих випадків.

Але не все так ідеально, цей метод також має свої недоліки. Наприклад, якщо одна точка є сусідньою до двох точок з різних кластерів, то вона доєднається до будь-якого з них, залежно від порядку обробки даних. Також до недоліків можна віднести те, що кластеризація та її якість залежать від функції відстані методу, а ще його нерентабельно використовувати на даних із великою щільністю.

РОЗДІЛ 2. ПРАКТИЧНА ЧАСТИНА

Для задачі сегментації ринку споживачів використовується персональний комп'ютер з процесором Apple M1.

Препроцесінг даних, візуалізація та кластеризація проводилась у середовищі PyCharm 2022.2, мова програмування – Python 3.10. Для цих задач використовувались такі бібліотеки: pandas, numpy, sklearn.cluster, sklearn.preprocessing, sklearn.metrics, seaborn, matplotlib.pyplot, kneed, scipy, sklearn.decomposition.

1 ПОСТАНОВКА ЗАДАЧІ

Розглянемо датасет, що представляє собою дані про 2206 відвідувачів певної компанії, що займається продажем товарів харчування. У датасеті містяться дані про профіль споживача, його переваги по товарах, його реакції на конкретні маркетингові кампанії та успішність різних каналів зв'язку із споживачем. Задача полягає у тому, щоб якісно провести сегментацію ринку споживачів цієї компанії, аби використовувати цю інформацію для подальшого розвитку маркетингу.

1.1 Огляд даних

У датасеті є інформація про 2206 клієнтів певного магазину, чи мережі магазинів продуктів харчування. Про кожного клієнта можемо знайти набір із 37-и показників. Опишемо детальніше кожну метрику:

1. Income – річний дохід споживача.
2. Kidhome – кількість малих дітей, що проживає в будинку в споживача.

3. Teenhome – кількість підлітків, що проживає в будинку споживача.
4. Recency – кількість днів від останньої покупки споживача.
5. MntWines – скільки грошей витрачено на товари категорії «Вино» споживачем за останніх два роки.
6. MntFruits – скільки грошей витрачено на товари категорії «Фрукти» споживачем за останніх два роки.
7. MntMeatProducts – скільки грошей витрачено на м’ясні товари споживачем за останніх два роки.
8. MntFishProducts – скільки грошей витрачено на рибні товари споживачем за останніх два роки.
9. MntSweetProducts – скільки грошей витрачено на товари категорії «Солодощі» споживачем за останніх два роки.
10. MntGoldProducts – скільки грошей витрачено на товари з золота споживачем за останніх два роки.
11. MntRegularProds – скільки грошей витрачено на звичайні товари з інших категорій.
12. NumDealsPurchases – кількість покупок, зроблених споживачем з використанням дисконтної картки.
13. NumWebPurchases – кількість покупок, зроблених споживачем через веб-сайт компанії.
14. NumCatalogPurchases – кількість покупок, зроблених споживачем через каталог товарів.
15. NumStorePurchases – кількість покупок, зроблених споживачем у магазині.
16. NumWebVisitsMonth – кількість відвідувань веб-сайту компанії споживачем за останній місяць.
17. AcceptedCmp1/2/3/4/5 – 1, якщо споживач прийняв офер в кампанії 1/2/3/4/5, інакше – 0. Для кожної кампанії окрема колонка.
18. AcceptedCmpOverall – скільки загалом прийнято оферів споживачем з різних кампаній.

19. Complain – 1, якщо споживач жалівся протягом останніх двох років, інакше – 0.
20. Response – 1, якщо споживач прийняв останній офер, інакше – 0.
21. Age – вік споживача.
22. Customer_Days – скільки днів пройшло від першої покупки споживача в цьому магазині.
23. marital_Divorced (Married, Single, Together, Widow) – сімейний статус споживача. 1 у відповідному стовпці, якщо у споживача саме цей сімейний статус.
24. education_2n Cycle (Basic, Graduation, Master, PhD) – рівень освіти споживача. 1 у відповідному стовпці, якщо у споживача саме цей рівень освіти.
25. MntTotal – скільки загалом витрачено грошей на покупки споживачем.

Тепер для кожної з наявних величин знайдемо її середнє значення, середньоквадратичне відхилення, мінімум, максимум та її квартилі. Результати на Рисунку 1.1

	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds
count	2205.000000	2205.000000	2205.000000	2205.000000	2205.000000	2205.000000	2205.000000	2205.000000	2205.000000	2205.000000
mean	51622.094785	0.442177	0.506576	49.009070	306.164626	26.403175	165.312018	37.756463	27.128345	44.057143
std	20713.063826	0.537132	0.544380	28.932111	337.493839	39.784484	217.784507	54.824635	41.130468	51.736211
min	1730.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	35196.000000	0.000000	0.000000	24.000000	24.000000	2.000000	16.000000	3.000000	1.000000	9.000000
50%	51287.000000	0.000000	0.000000	49.000000	178.000000	8.000000	68.000000	12.000000	8.000000	25.000000
75%	68281.000000	1.000000	1.000000	74.000000	507.000000	33.000000	232.000000	50.000000	34.000000	56.000000
max	113734.000000	2.000000	2.000000	99.000000	1493.000000	199.000000	1725.000000	259.000000	262.000000	321.000000

	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1
count	2205.000000	2205.000000	2205.000000	2205.000000	2205.000000	2205.000000	2205.000000	2205.000000	2205.000000
mean	2.318367	4.100680	2.645351	5.823583	5.336961	0.073923	0.074376	0.073016	0.064399
std	1.886107	2.737424	2.798647	3.241796	2.413535	0.261705	0.262442	0.260222	0.245518
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	2.000000	0.000000	3.000000	3.000000	0.000000	0.000000	0.000000	0.000000
50%	2.000000	4.000000	2.000000	5.000000	6.000000	0.000000	0.000000	0.000000	0.000000
75%	3.000000	6.000000	4.000000	8.000000	7.000000	0.000000	0.000000	0.000000	0.000000
max	15.000000	27.000000	28.000000	13.000000	20.000000	1.000000	1.000000	1.000000	1.000000

	AcceptedCmp2	Complain	Response	Age	Customer_Days	marital_Divorced	marital_Married	marital_Single	marital_Together	marital_Widow
count	2205.000000	2205.000000	2205.00000	2205.000000	2205.000000	2205.000000	2205.000000	2205.000000	2205.000000	2205.000000
mean	0.013605	0.009070	0.15102	51.095692	2512.718367	0.104308	0.387302	0.216327	0.257596	0.034467
std	0.115872	0.094827	0.35815	11.705801	202.563647	0.305730	0.487244	0.411833	0.437410	0.182467
min	0.000000	0.000000	0.00000	24.000000	2159.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.00000	43.000000	2339.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.00000	50.000000	2515.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	0.000000	0.000000	0.00000	61.000000	2688.000000	0.000000	1.000000	0.000000	1.000000	0.000000
max	1.000000	1.000000	1.00000	80.000000	2858.000000	1.000000	1.000000	1.000000	1.000000	1.000000

	education_2n Cycle	education_Basic	education_Graduation	education_Master	education_PhD	MntTotal	MntRegularProds	AcceptedCmpOverall
count	2205.000000	2205.000000	2205.000000	2205.000000	2205.000000	2205.000000	2205.000000	2205.00000
mean	0.089796	0.024490	0.504762	0.165079	0.215873	562.764626	518.707483	0.29932
std	0.285954	0.154599	0.500091	0.371336	0.411520	575.936911	553.847248	0.68044
min	0.000000	0.000000	0.000000	0.000000	0.000000	4.000000	-283.000000	0.00000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	56.000000	42.000000	0.00000
50%	0.000000	0.000000	1.000000	0.000000	0.000000	343.000000	288.000000	0.00000
75%	0.000000	0.000000	1.000000	0.000000	0.000000	964.000000	884.000000	0.00000
max	1.000000	1.000000	1.000000	1.000000	1.000000	2491.000000	2458.000000	4.00000

Рисунок 1.1 - Опис кожної метрики з датасету

Тепер проглянемо датасет на наявність null – значень або «ненормальних» даних. Можемо помітити, що у колонці MntRegularProds мінімальне значення від'ємне, чого на практиці не може бути. Споживачів з від'ємними значеннями цього показника всього троє, тому просто виключимо їх з датасету.

На рисунку 1.2 можемо побачити, що null – значень у датасеті немає.

```
0   Income                2202 non-null   float64
1   Kidhome               2202 non-null   int64
2   Teenhome              2202 non-null   int64
3   Recency               2202 non-null   int64
4   MntWines              2202 non-null   int64
5   MntFruits             2202 non-null   int64
6   MntMeatProducts       2202 non-null   int64
7   MntFishProducts       2202 non-null   int64
8   MntSweetProducts      2202 non-null   int64
9   MntGoldProds          2202 non-null   int64
10  NumDealsPurchases     2202 non-null   int64
11  NumWebPurchases       2202 non-null   int64
12  NumCatalogPurchases   2202 non-null   int64
13  NumStorePurchases     2202 non-null   int64
14  NumWebVisitsMonth     2202 non-null   int64
15  AcceptedCmp3          2202 non-null   int64
16  AcceptedCmp4          2202 non-null   int64
17  AcceptedCmp5          2202 non-null   int64
18  AcceptedCmp1          2202 non-null   int64
19  AcceptedCmp2          2202 non-null   int64
20  Complain              2202 non-null   int64
21  Response              2202 non-null   int64
22  Age                   2202 non-null   int64
23  Customer_Days         2202 non-null   int64
24  marital_Divorced      2202 non-null   int64
25  marital_Married       2202 non-null   int64
26  marital_Single        2202 non-null   int64
27  marital_Together      2202 non-null   int64
28  marital_Widow         2202 non-null   int64
29  education_2n Cycle    2202 non-null   int64
30  education_Basic       2202 non-null   int64
31  education_Graduation  2202 non-null   int64
32  education_Master      2202 non-null   int64
33  education_PhD         2202 non-null   int64
34  MntTotal              2202 non-null   int64
35  MntRegularProds       2202 non-null   int64
36  AcceptedCmpOverall    2202 non-null   int64
```

Рисунок 1.2 - Інформація про колонки датасету

Для кращого розуміння даних, додаємо гістограми основних метрик (рис. 1.3):

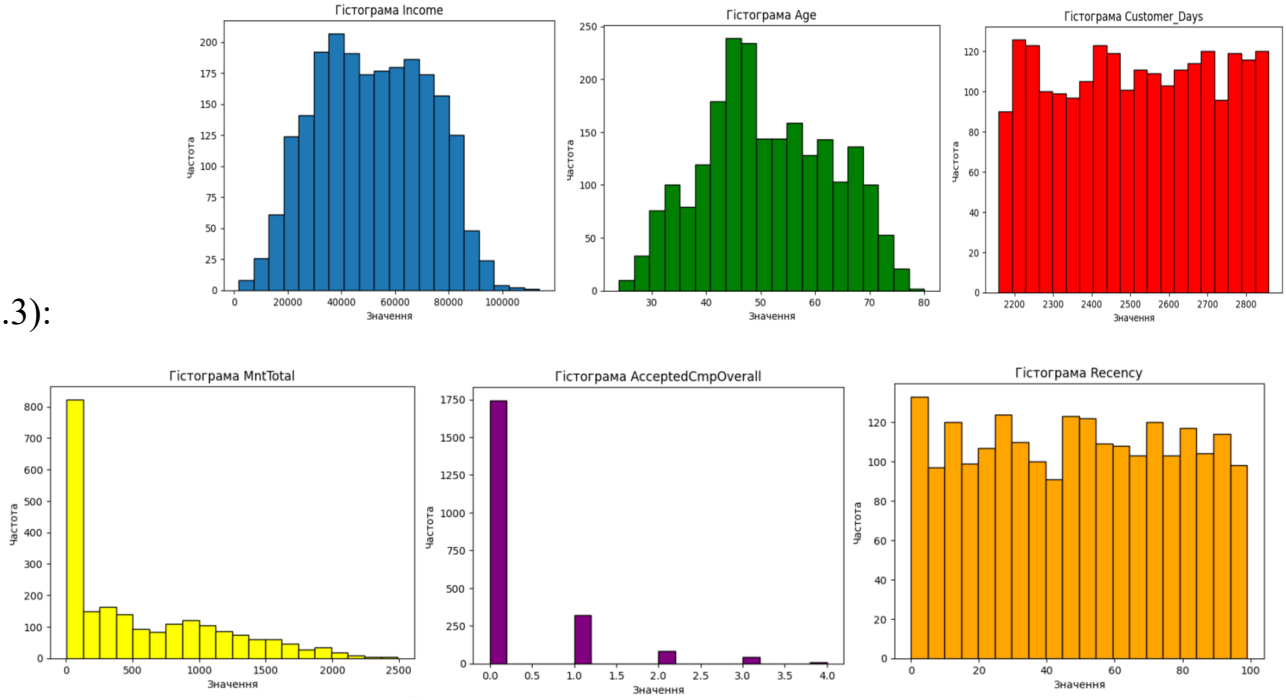

Рисунок 1.3 - Гістограми деяких метрик датасету

Ще одним етапом дослідження датасету буде побудова кореляційної матриці, щоб зрозуміти, чи багато залежних змінних у датасеті. Кореляційна матриця зображена на рисунку 1.4

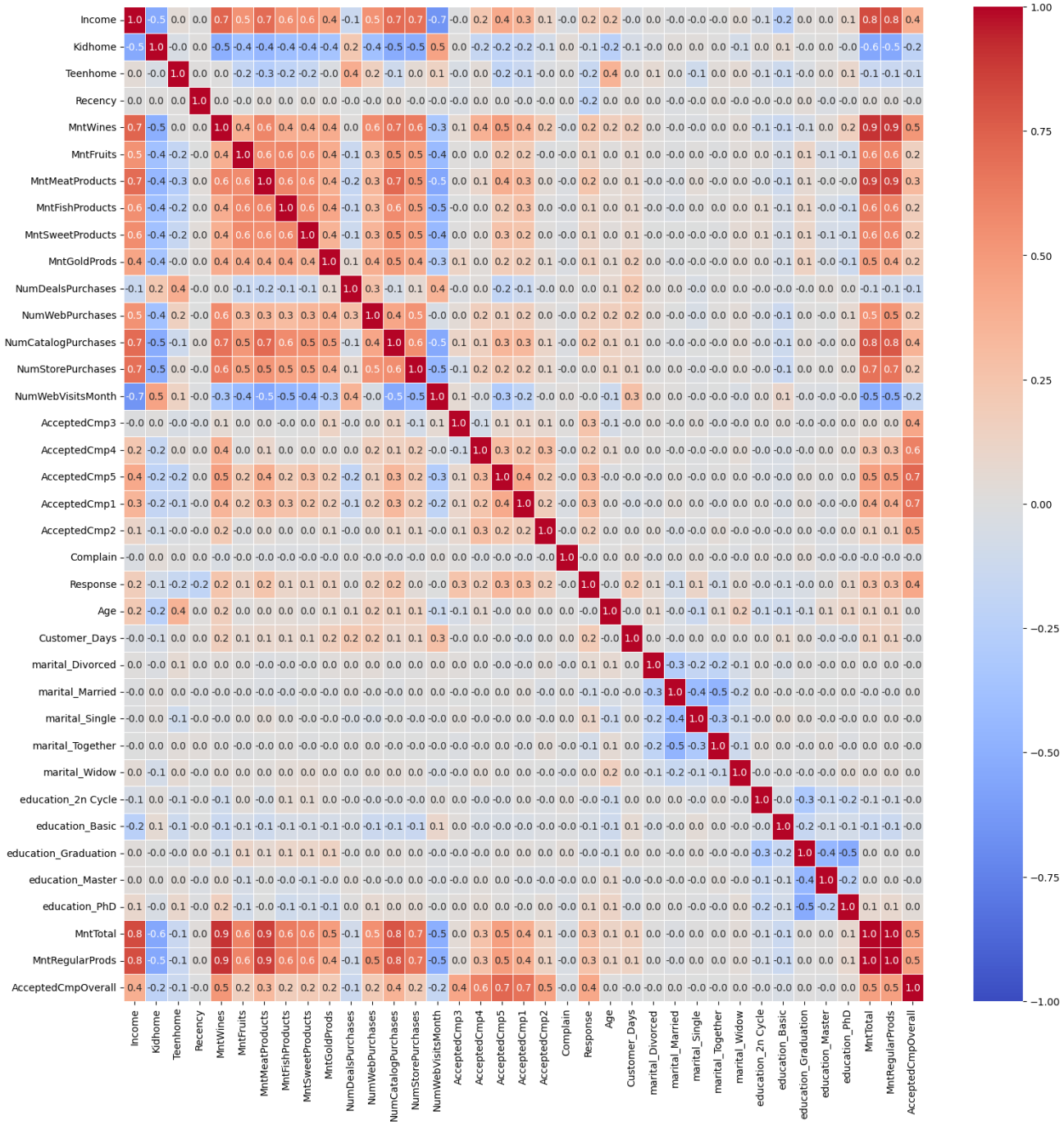

Рисунок 1.4 - Кореляційна матриця датасету

За цією кореляційною матрицею можемо побачити, що у нас є метрики, які мають високу кореляцію, а також метрики, які ні з чим не корелюють.

1.2 Препроцесінг даних

Щоб спростити задачу кластеризації, можемо зменшити розмірність даних та позбутись «схожих» або «некорисних» метрик. Спочатку видалимо колонки, у яких немає кореляції з іншими: Recency, Complain, marital, education, MntTotal. Далі понизимо розмірність, для цього використаємо метод головних компонент, або PCA (principal component analysis), та метод ліктя для інтерпретації даних, перед цим стандартизувавши дані методом Z-Scores. Результат методу представлено на рис. 1.5

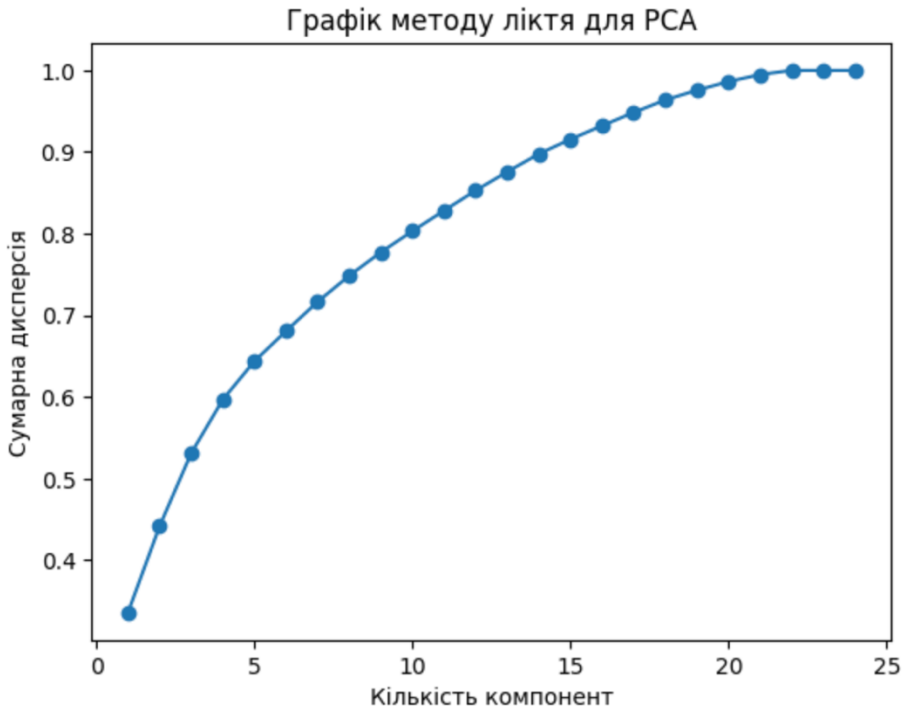

Рисунок 1.5 - Результат методу PCA

На графіку видно, що крива перестає збільшуватись після 21-ої компоненти. 21 метрика покриває 99.9999% дисперсії. При цьому, 14 компонент пояснюють 89.6% дисперсії, тому зупинимось на цьому числі. Проводимо метод головних компонент на нашому датасеті для 14 компонент. Результат на рис. 1.6

```
Пояснена дисперсія кожної компоненти: [0.33406814 0.10583804 0.08876204 0.0655764  0.04716563 0.03684957
 0.03601937 0.03184391 0.02947292 0.02601475 0.02519027 0.02448072
 0.02296222 0.02208678]
Загальна пояснена дисперсія: 0.8963307667070194
```

Рисунок 1.6 - Результат методу головних компонент

Ще перед проведенням кластеризації бажано візуалізувати дані. Так як у нас тепер 14-вимірних простір, доведеться знову звернутись до методів пониження розмірності, аби хоч приблизно розуміти, що з себе представляють дані та для їх кращого сприйняття. Результат зображено на рис. 1.7

Рисунок 1.7 - Зображення датасету у 2-вимірному просторі

2 КЛАСТЕРИЗАЦІЯ

Загалом, з рисунку даних не зовсім зрозуміло, який метод краще підійде для кластеризації, тому проведемо її декількома методами, та порівняємо результати.

Перевірятимемо якість кластеризацій будемо за допомогою індекса Davies-Bouldin та Silhouette score. Нижче значення метрики Davies-Bouldin вказує на вищу якість кластеризації: добру відокремленість кластерів та мінімальну перекритість між ними, а в Silhouette score навпаки - вище значення вказує на кращу кластеризацію. Індекс Davies-Bouldin знаходиться як середнє значення максимального відношення між відстанню точки від центра її групи і відстанню між двома центроїдами.

2.1 Метод K-MEANS

Спочатку спробуємо визначити, яка кількість кластерів є оптимальною для кластеризації. Перевіримо це двома методами: методом ліктя та методом силуету.

Для методу ліктя проведемо кластеризацію методом K-Means для кількості кластерів від 2 до 100. Для ініціалізації центроїдів використаємо метод k-means++. Функція відстані – Евклідова.

За методом ліктя, маємо такий результат: рис. 2.1

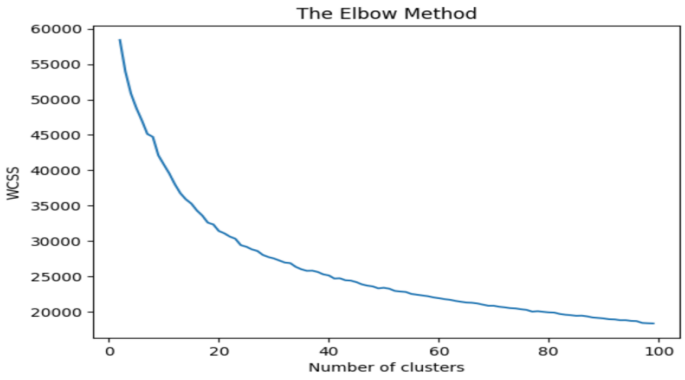

Рисунок 2.1 - Метод ліктя

Оптимальна кількість кластерів – 26.

Тепер перевіримо те саме, але методом силуету. Метод силуету показує, що оптимальна кількість кластерів – 2, при цьому коефіцієнт силуету дорівнює 0.31. Як бачимо, наші результати зовсім не сходяться. Зобразимо кластеризацію для 26 та 2 кластерів методом k-середніх та порівняємо. На рис. 2.2 зображення методу силуету для двох кластерів:

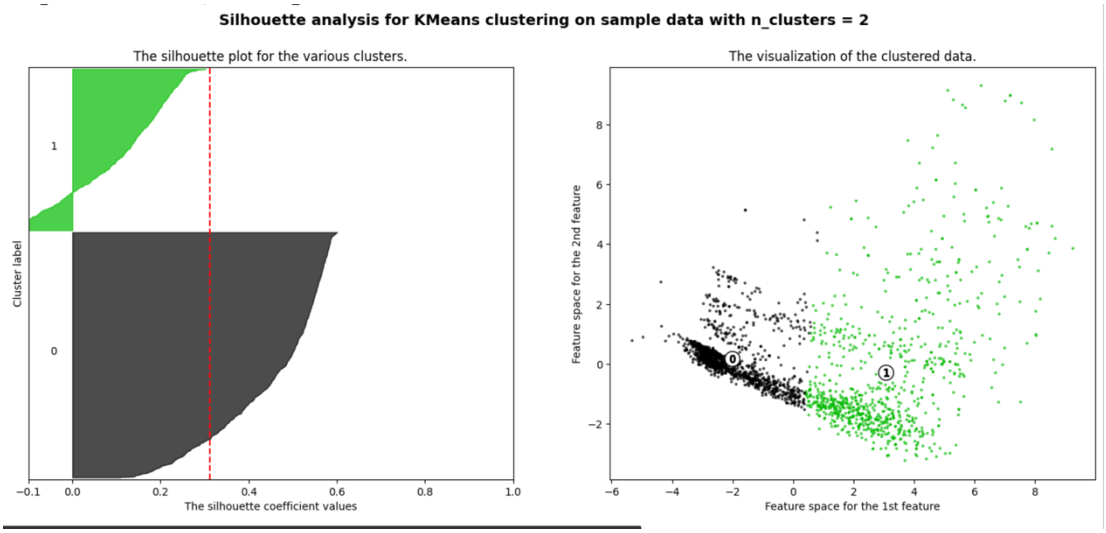

Рисунок 2.2 - Метод силуету для двох кластерів

Результат методу k-середніх для 26 кластерів зображено на рис. 2.3

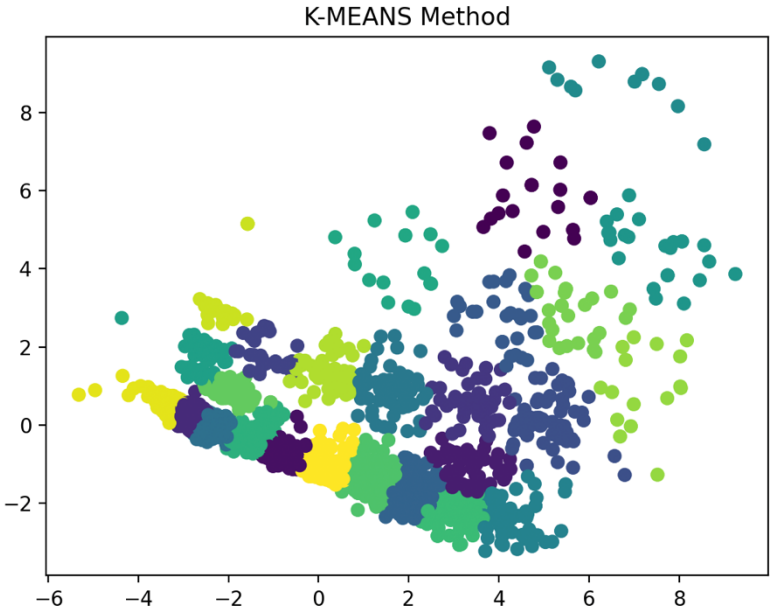

Рисунок 2.3 - Кластеризація методом k-середніх

2.2 Метод Варда

Тепер спробуємо провести кластеризацію ієрархічним методом Варда. Спершу побудуємо дендрограму: рис. 2.3

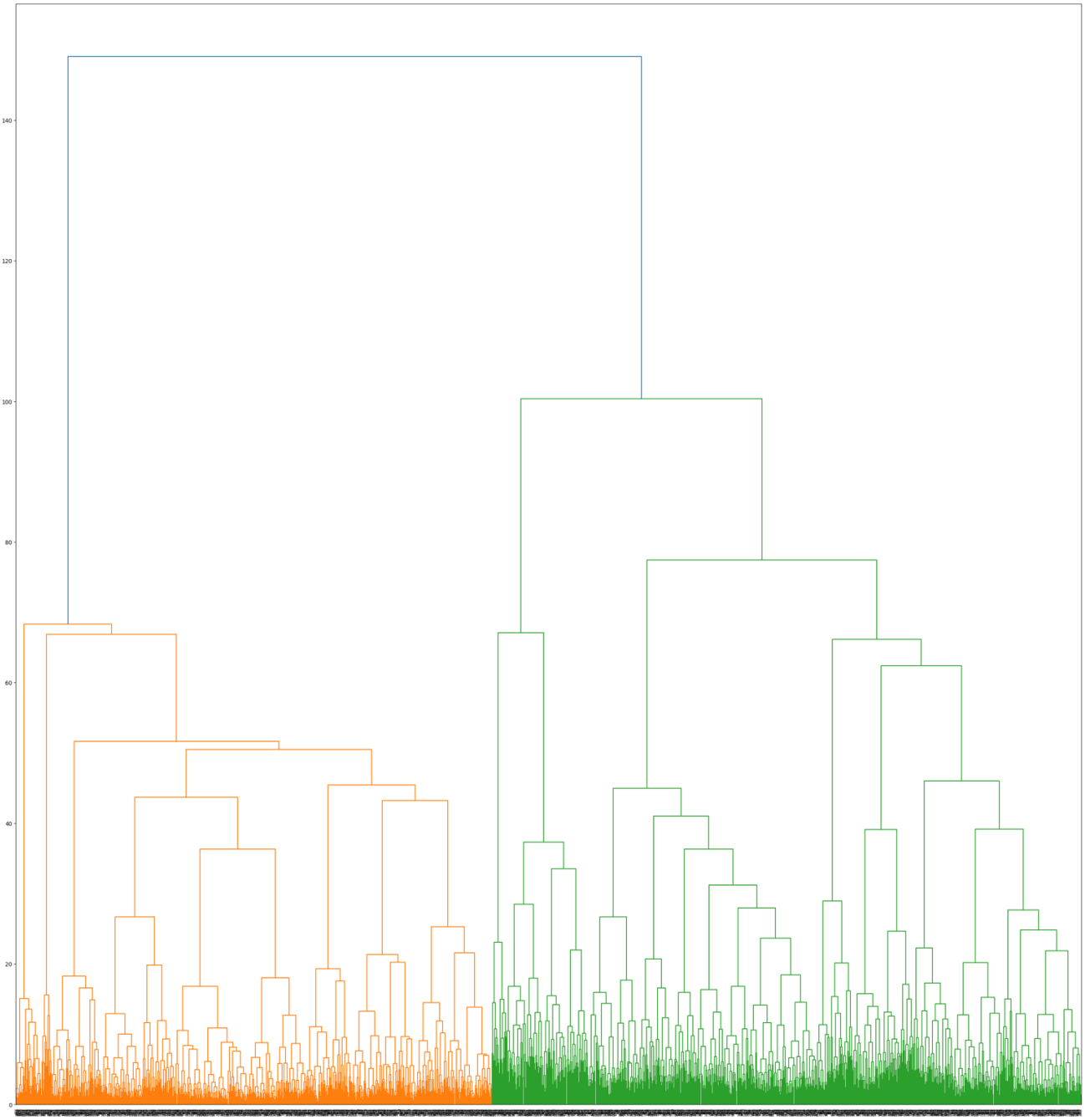

Рисунок 2.3 - Дендрограма

Знайдемо оптимальну кількість кластерів для Методу Варда. Для цього ми скористаємось методом ліктя на даних із дендрограми. Результат показує, що оптимальна кількість кластерів – два. Проводимо кластеризацію для наших даних

використовуючи метод Варда. У параметрах методу вказуємо кількість кластерів – 2, функцію відстані – Евклідова. Зобразимо наш результат: рис. 2.4

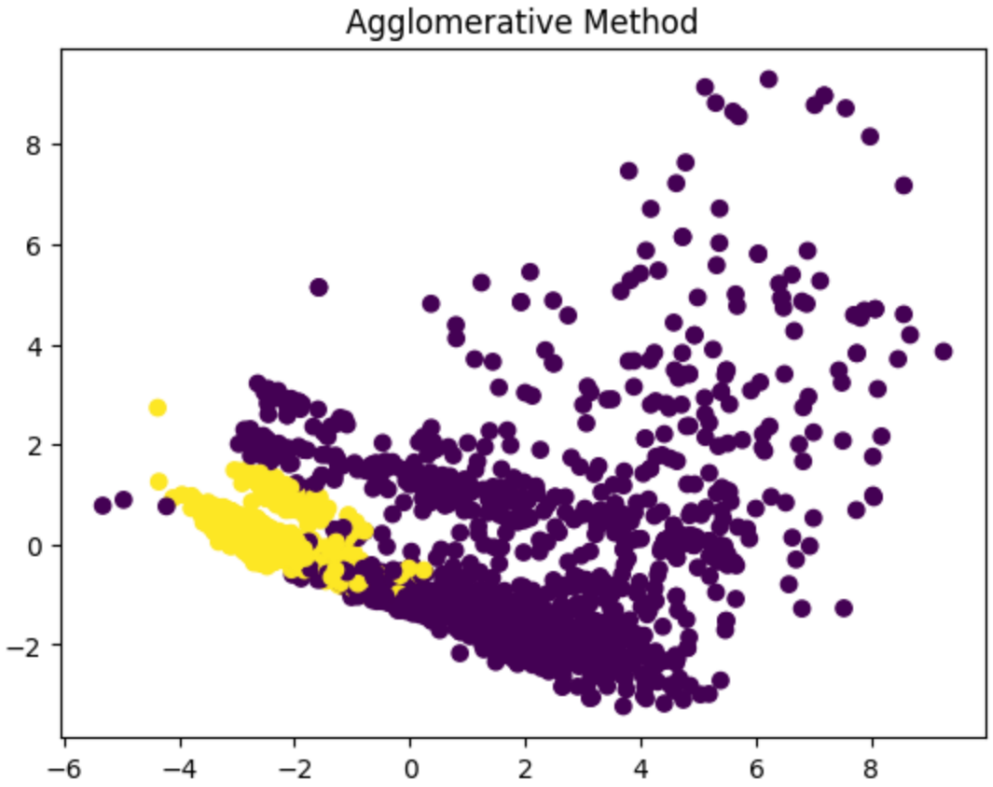

Рисунок 2.4 - Кластеризація методом Варда

2.3 Метод DBSCAN

Останній метод, яким буде проводитись кластеризація – метод, що заснований на щільності точок – DBSCAN. Для того, щоб знайти оптимальний параметр eps, пройдемось циклом для всіх значень eps від 0.001 до 0.5 з кроком 0.001. Те саме зробимо для параметра minPts. Кожне значення обох параметрів підставляємо у метод кластеризації. Результат показав, що оптимальні значення параметрів такі: eps = 0.01, minPts = 2.

З цими параметрами проводимо кластеризацію. Функція відстані – Евклідова.

Метод DBSCAN розділив початковий датасет на 179 кластерів, де в першому кластерів зібрано 1840 споживачів. Кластеризація має такий вигляд: рис. 2.5

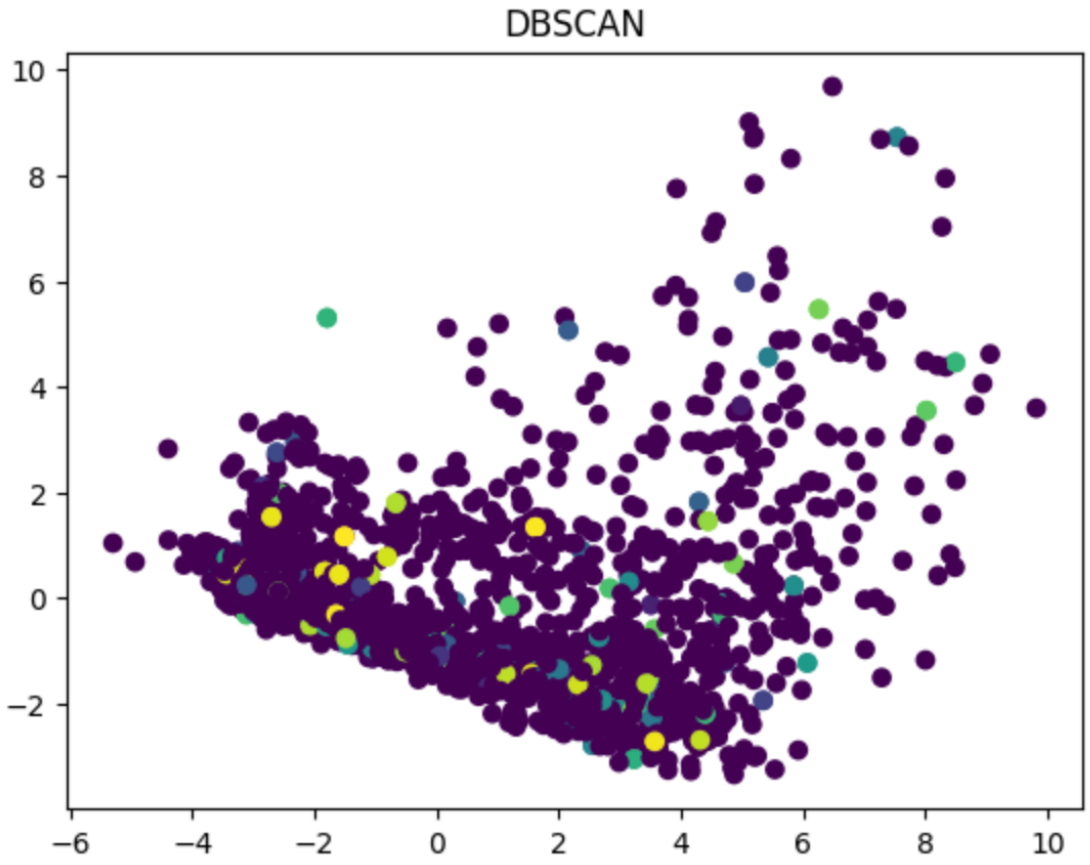

Рисунок 2.5 - Кластеризація методом DBSCAN

3 РЕЗУЛЬТАТ

Для порівняння результатів усіх методів кластеризації, подивимось на значення індекса Davies-Bouldin та Silhouette score у таблиці 3.1

Таблиця 3.1 - Порівняння методів кластеризації

метод	кількість кластерів	Davies-Bouldin	Silhouette
K-Means	26	3.45	0.005
K-Means	2	1.46	0.31
Ward	2	1.56	0.22
DBSCAN	179	1.14	-0.4

З таблиці 3.1 видно, що найкращими є методи K-Means з двома кластерами та метод Варда. Інші методи не є задовільними як через показники метрик якості, так і з певних логічних міркувань, адже ділити ринок на 26 чи 179 сегментів не дуже вигідно, бо придумати стратегії для кожного з цих кластерів буде дуже дорого, або обираючи один з них для просування відкидається занадто велика частина споживачів.

Отож, розглянемо детальніше результати кластеризації методами k-середніх та Варда із поділом на два кластери.

1. Метод K-Means.

0 кластер – 1328 об’єктів, 1 кластер – 874 об’єкти

clusters_kmeans	Income	Kidhome	Teenhome	Recency	MntWines
0	38643.301958	0.692018	0.557229	48.889307	99.953313
1	71499.178490	0.064073	0.427918	49.251716	620.355835

clusters_kmeans	MntFruits	MntMeatProducts	MntFishProducts
0	7.011295	37.920934	10.246235
1	55.937071	359.362700	79.673913

clusters_kmeans	MntSweetProducts	MntGoldProds	NumDealsPurchases
0	7.140813	22.915663	2.528614
1	57.581236	75.331808	2.006865

clusters_kmeans	NumWebPurchases	NumCatalogPurchases	NumStorePurchases
0	2.987199	0.897590	3.975151
1	5.750572	5.308924	8.649886

clusters_kmeans	AcceptedCmp1	AcceptedCmp2	Complain	Response	Age
0	0.005271	0.001506	0.009789	0.094127	49.785392
1	0.154462	0.032037	0.008009	0.237986	53.098398

clusters_kmeans	Customer_Days	marital_Divorced	marital_Married
0	2497.283133	0.096386	0.399849
1	2536.618993	0.116705	0.368421

clusters_kmeans	marital_Single	marital_Together	marital_Widow
0	0.216114	0.259789	0.027861
1	0.215103	0.255149	0.044622

clusters_kmeans	education_2n Cycle	education_Basic	education_Graduation
0	0.095633	0.039910	0.489458
1	0.081236	0.001144	0.528604

clusters_kmeans	education_Master	education_PhD	MntTotal
0	0.173193	0.201807	162.272590
1	0.152174	0.236842	1172.910755

clusters_kmeans	MntRegularProds	AcceptedCmpOverall	clusters_ward
0	139.356928	0.103163	0.700301
1	1097.578947	0.598398	0.000000

0 кластер: менший рівень доходу, менші витрати на покупки, менше реагують на рекламні кампанії, менше роблять покупок, але більше заходять на веб-сайт магазину.

1 кластер: протилежний до 0 кластера.

2. Метод Варда.

0 кластер – 1272 об’єкти, 1 кластер – 930 об’єктів

clusters_ward	Income	Kidhome	Teenhome	Recency	MntWines
0	64074.865566	0.198899	0.550314	48.878931	499.463836
1	34736.944086	0.776344	0.445161	49.244086	42.592473

clusters_ward	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts
0	41.876572	268.747642	59.825472	43.164308
1	5.304301	24.295699	7.681720	5.273118

clusters_ward	MntGoldProds	NumDealsPurchases	NumWebPurchases
0	65.320755	2.576258	5.513365
1	14.176344	1.973118	2.129032

clusters_ward	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth
0	4.206761	7.654088	4.592767
1	0.517204	3.336559	6.352688

clusters_ward	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1
0	0.127358	0.128931	0.126572	0.111635
1	0.001075	0.000000	0.000000	0.000000

clusters_ward	AcceptedCmp2	Complain	Response	Age	Customer_Days
0	0.023585	0.007075	0.205189	52.900943	2538.726415
1	0.000000	0.011828	0.077419	48.637634	2477.566667

clusters_ward	marital_Divorced	marital_Married	marital_Single
0	0.110063	0.383648	0.208333
1	0.096774	0.392473	0.225806

clusters_ward	marital_Together	marital_Widow	education_2n Cycle	education_Basic	education_Graduation	education_Master	education_PhD	MntTotal	MntRegularProds	AcceptedCmpOverall
0	0.257075	0.040881	0.076258	0.006289	0.515723	0.164308	0.237421	913.077830	847.757075	0.518082
1	0.259140	0.025806	0.108602	0.049462	0.490323	0.165591	0.186022	85.147312	70.970968	0.001075

0 кластер: більший рівень доходу, більші витрати на продукти, здійснюють більше покупок, більше реагують на рекламні пропозиції.

1 кластер: протилежний до 0 кластера.

Як бачимо, обоє методів виділили в окремий кластер споживачів із вищим рівнем доходу, які більше витрачають грошей на покупки та більше реагують на рекламні кампанії. Проте у методі K-Means ця група людей дещо менша, але з більшим середнім доходом, кращою реакцією на рекламні пропозиції та більшими витратами на покупки. Спрямування стратегії розвитку такого сегменту принесе компанії більший прибуток.

Отже, за двома метриками якості кластеризації, а також за певними логічними міркуваннями, найкращий результат виявився у метода K-Means із поділом на два кластери. Цей алгоритм розділив ринок на дві групи, одна з яких є гарним кандидатом на застосування до неї маркетингової стратегії: високий рівень доходу, гарна реакція на рекламні кампанії, велика частота та витрати на покупки, методом взаємодії із цією групою будуть каталог та фізичний магазин.

ВИСНОВКИ

У роботі були досліджені різні методи кластерного аналізу задля вирішення задачі сегментації ринку, розглянуті етапи вирішення цієї задачі та різні методи.

Теоретично було розглянуто ієрархічні (метод Варда, метод найближчого сусіда, метод середніх), та неієрархічні (метод K-Means та DBSCAN) методи кластеризації. Неієрархічні та метод Варда були практично реалізовані та візуалізовані за допомогою мови програмування Python. Результати методів порівнювались за допомогою індекса Davies-Bouldin та Silhouette score.

Результати роботи трьох алгоритмів кластеризації для одного набору даних дозволяють зробити наступні висновки:

1. Потрібно перевіряти декількома методами оптимальну кількість кластерів.
2. Окрім заданих алгоритмів знаходження оптимальної кількості кластерів, необхідно ще опиратись на логічні міркування щодо кількості сегментів споживачів.
3. Якість кластеризації краще перевіряти декількома метриками.
4. Метод DBSCAN не завжди якісно кластеризує дані, коли у нас багато об’єктів з мінімальними відмінностями, що створює обширну зону із великою щільністю.
5. Методи K-MEANS та метод Варда можуть давати дуже схожі результати для однакової кількості кластерів із однаковою заданою функцією відстані.
6. Метод K-MEANS найкраще відокремив кластер споживачів, на яку можна зосередити більше уваги і очікувати найефективнішого результату.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Kotler, P., & Keller, K. L. (2012). Marketing management. Pearson Education Limited.
2. Armstrong, G., & Kotler, P. (2017). Marketing: an introduction. Pearson.
3. "Principles of Marketing" by Philip Kotler and Gary Armstrong
4. "Marketing Management" by Philip Kotler and Kevin Lane Keller
5. "Market Segmentation Techniques" by Barry J. Babin and William G. Zikmund
6. Гаркавенко С.С. Маркетинг: Підручник. – 5-те вид. доп. – Київ: Лібра, 2007.
7. Крамаренко В.І. Маркетинг: Навчальний посібник. – К.: ЦУЛ, 2003
8. Балабанова Л.В. Маркетинг. – Донецьк, 2002.
9. "Cluster Analysis: Algorithms for Data Reduction and Classification of Objects" by J. A. Hartigan and M. A. Wong
10. "Cluster Analysis for Data Mining and System Identification" by Guanrong Chen

ДОДАТОК А. Код програми

```
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
import seaborn as sns
import matplotlib.pyplot as plt
from kneed import KneeLocator
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.metrics import davies_bouldin_score
from sklearn import preprocessing
from scipy.stats import iqr
from scipy import stats
import matplotlib.cm as cm
from sklearn.metrics import silhouette_samples, silhouette_score
import scipy.cluster.hierarchy as sch
from sklearn.cluster import AgglomerativeClustering
from sklearn.cluster import DBSCAN


filepath = 'ifood_df.csv'
df = pd.read_csv(filepath, sep=',')
pd.set_option('display.max_columns', None)


# check for null-values
print("Rows:", df.shape[0])
print("Features:", df.shape[1])
print()
print(df.info())

# describe every column
print(df.describe())

df = df.drop(labels=['Z_CostContact', 'Z_Revenue'], axis=1)
df = df[df.MntRegularProds >= 0]

# hist for some metrics
plt.hist(df['Recency'], bins=20, edgecolor='black', color='orange')
plt.xlabel('Значення')
plt.ylabel('Частота')
```

```
plt.title('Гістограма Recency')
plt.show()

# standarting data
sc = StandardScaler()
df1 = pd.DataFrame(sc.fit_transform(df.iloc[:, :].values), columns=df.columns)
print(df1.describe())

# correlation map
fig, ax = plt.subplots(figsize=(18, 18))
sns.heatmap(df1.corr(), vmin=-1, vmax=+1, annot=True, cmap='coolwarm',
linewidths=.5, ax=ax, fmt='.1f')
plt.show()


# drop some columns with zero correlation
df1 = df1[['Income', 'Kidhome', 'Teenhome', 'MntWines', 'MntFruits',
       'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts',
       'MntGoldProds', 'NumDealsPurchases', 'NumWebPurchases',
       'NumCatalogPurchases', 'NumStorePurchases', 'NumWebVisitsMonth',
       'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', 'AcceptedCmp1',
       'AcceptedCmp2', 'Response',
       'Age', 'Customer_Days', 'MntRegularProds',
       'AcceptedCmpOverall']]

# PCA
max_components = min(df1.shape[0], df1.shape[1])
num_components = range(1, max_components + 1)
variances = []
for n in num_components:
    pca = PCA(n_components=n)
    pca.fit(df1)
    variance = pca.explained_variance_ratio_.sum()
    variances.append(variance)

plt.plot(num_components, variances, marker='o')
plt.xlabel('Кількість компонент')
plt.ylabel('Сумарна дисперсія')
plt.title('Графік методу ліктя для PCA')
plt.show()

n_components = 14

pca = PCA(n_components=n_components)
pca_result = pca.fit_transform(df1)
```

```
explained_variance_ratio = pca.explained_variance_ratio_
print('Пояснена дисперсія кожної компоненти:', explained_variance_ratio)

total_variance_explained = sum(explained_variance_ratio)
print('Загальна пояснена дисперсія:', total_variance_explained)


# visualisation with PCA
pca = PCA(n_components=2)
reduced_data = pca.fit_transform(pca_result)

plt.scatter(reduced_data[:, 0], reduced_data[:, 1], c='blue', alpha=0.5)
plt.xlabel('Головна компонента 1')
plt.ylabel('Головна компонента 2')
plt.title('Діаграма розсіювання після PCA')
plt.show()


# clustering with k-means method
X = pca_result
wcss = []
best_score = -1
best_n_clusters = 0
for i in range(2, 100):
    kmeans = KMeans(n_clusters=i, init='k-means++', random_state=42)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)
    labels = kmeans.labels_
    score = silhouette_score(X, labels)
    if score > best_score:
        best_score = score
        best_n_clusters = i
plt.plot(range(2, 100), wcss)
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()

kl = KneeLocator(
    range(2,  100), wcss, curve="convex", direction="decreasing"
)

print(kl.elbow)
print(best_n_clusters, best_score)
```

```
# visualisation


fig, (ax1, ax2) = plt.subplots(1, 2)
fig.set_size_inches(18, 7)
ax1.set_xlim([-0.1, 1])
ax1.set_ylim([0, len(X) + (2 + 1) * 10])

clusterer = KMeans(n_clusters = 2, init = 'k-means++', random_state = 42)
cluster_labels = clusterer.fit_predict(X)

silhouette_avg = silhouette_score(X, cluster_labels)
print(
    "For n_clusters =",
    2,
    "The average silhouette_score is :",
    silhouette_avg,
)

sample_silhouette_values = silhouette_samples(X, cluster_labels)

y_lower = 10
for i in range(2):
    ith_cluster_silhouette_values = sample_silhouette_values[cluster_labels == i]

    ith_cluster_silhouette_values.sort()

    size_cluster_i = ith_cluster_silhouette_values.shape[0]
    y_upper = y_lower + size_cluster_i

    color = cm.nipy_spectral(float(i) / 2)
    ax1.fill_betweenx(
        np.arange(y_lower, y_upper),
        0,
        ith_cluster_silhouette_values,
        facecolor=color,
        edgecolor=color,
        alpha=0.7,
    )

    ax1.text(-0.05, y_lower + 0.5 * size_cluster_i, str(i))

    y_lower = y_upper + 10
```

```
    ax1.set_title("The silhouette plot for the various clusters.")
    ax1.set_xlabel("The silhouette coefficient values")
    ax1.set_ylabel("Cluster label")

    ax1.axvline(x=silhouette_avg, color="red", linestyle="--")

    ax1.set_yticks([])
    ax1.set_xticks([-0.1, 0, 0.2, 0.4, 0.6, 0.8, 1])

colors = cm.nipy_spectral(cluster_labels.astype(float) / 2)
ax2.scatter(
        X[:, 0], X[:, 1], marker=".", s=30, lw=0, alpha=0.7, c=colors, edgecolor="k"
    )

centers = clusterer.cluster_centers_

ax2.scatter(
    centers[:, 0],
    centers[:, 1],
    marker="o",
    c="white",
    alpha=1,
    s=200,
    edgecolor="k",
)

for i, c in enumerate(centers):
    ax2.scatter(c[0], c[1], marker="$%d$" % i, alpha=1, s=50, edgecolor="k")

ax2.set_title("The visualization of the clustered data.")
ax2.set_xlabel("Feature space for the 1st feature")
ax2.set_ylabel("Feature space for the 2nd feature")

plt.suptitle(
    "Silhouette analysis for KMeans clustering on sample data with n_clusters = %d"
     % 2,
    fontsize=14,
    fontweight="bold",
    )

plt.show()

best_kmeans = KMeans(n_clusters=26, init='k-means++')
best_kmeans.fit(reduced_data)
```

```
x = []
y = []
for i in reduced_data:
  x.append(i[0])
  y.append(i[1])
plt.scatter(x=x, y=y, c=best_kmeans.labels_)
plt.title('K-MEANS Method')
plt.show()

db_index_kmeans_elbow = davies_bouldin_score(X, best_kmeans.labels_)
db_index_kmeans_silhouette = davies_bouldin_score(X, cluster_labels)
silhouette = silhouette_score(X, cluster_labels)
silhouette_elbow = silhouette_score(X, best_kmeans.labels_)


# hierarchy method
linkage_matrix = sch.linkage(X, method='ward')

# Побудова дендрограми
plt.figure(figsize=(10, 7))
dendrogram = sch.dendrogram(linkage_matrix)
plt.xlabel('Об\'єкти')
plt.ylabel('Відстань')
plt.title('Дендрограма')

# Пошук оптимальної кількості кластерів за методом "еллбоу"
distances = linkage_matrix[:, 2]
deltas = distances[:-1] - distances[1:]
num_clusters = deltas.argmax() + 2  # +2, оскільки індексація з 0

plt.axvline(x=num_clusters - 0.5, color='red', linestyle='--', label='Оптимальна
кількість кластерів')
plt.legend()
plt.show()

print(f"Оптимальна кількість кластерів: {num_clusters}")

hc = AgglomerativeClustering(n_clusters=2, affinity='euclidean', linkage='ward')
y_hc = hc.fit_predict(X)
plt.scatter(x=x, y=y, c=y_hc)
plt.title('Agglomerative Method')
plt.show()

db_index_hierarchy = davies_bouldin_score(X, hc.labels_)
silhouette_hierarchy = silhouette_score(X, hc.labels_)
```

```
# dbscan

dbscan = DBSCAN(eps=0.01, min_samples=2)
clusters_dbscan = dbscan.fit_predict(X)
db_index_dbscan = davies_bouldin_score(X, dbscan.labels_)
silhouette_dbscan = silhouette_score(X, dbscan.labels_)

plt.scatter(x=x, y=y, c=clusters_dbscan)
plt.title('DBSCAN')
plt.show()
# result metrics
print('Davies Bouldin Index K-Means Elbow: ', db_index_kmeans_elbow)
print('Davies Bouldin Index K-Means Silhouette: ', db_index_kmeans_silhouette)
print('Davies Bouldin Index Ward: ', db_index_hierarchy)
print('Davies Bouldin Index DBSCAN: ', db_index_dbscan)
print('Silhouette score K-Means Elbow: ', silhouette_elbow)
print('Silhouette score K-Means Silhouette: ', silhouette)
print('Silhouette score Ward: ', silhouette_hierarchy)
print('Silhouette score DBSCAN: ', silhouette_dbscan)
df['clusters_kmeans'] = cluster_labels
df['clusters_ward'] = y_hc
print(df.groupby('clusters_kmeans').mean())
print(df.groupby('clusters_ward').mean())
```