МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ВАСИЛЯ СТУСА
ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ І ПРИКЛАДНИХ ТЕХНОЛОГІЙ
КАФЕДРА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

**НАЗАРЕНКО МАРІЯ СЕРГІЇВНА**

Допускається до захисту:
в.о. завідувача кафедри
інформаційних технологій,
д.т.н., професор
_____ Наталія ВЕСЕЛОВСЬКА
«____»_____ 2026 р.

# РОЗРОБКА АНАЛІТИЧНОЇ СИСТЕМИ СЕГМЕНТАЦІЇ КЛІЄНТІВ НА ОСНОВІ МЕТОДІВ СТАТИСТИЧНОГО НАВЧАННЯ ІЗ ЗАСТОСУВАННЯМ ІНСТРУМЕНТІВ БІЗНЕС-АНАЛІТИКИ

Спеціальність 122 Комп’ютерні науки

**Кваліфікаційна (бакалаврська) робота**

Керівник:
Юрій ПОРЕМСЬКИЙ, старший викладач кафедри інформаційних технологій, канд. техн. наук.
_____

Оцінка: _____ / _____ / _____
(бали/за шкалою ЄКТС/за національною шкалою)

Голова ЕК: _____
(підпис)

Вінниця – 2026

## АНОТАЦІЯ

**Назаренко М. С.** Розробка аналітичної системи сегментації клієнтів на основі методів статистичного навчання із застосуванням інструментів бізнес-аналітики. Спеціальність 122 «Комп'ютерні науки». Донецький національний університет імені Василя Стуса, Вінниця, 2026.

У кваліфікаційній роботі розглянуто задачу сегментації клієнтів на основі даних про їхню купівельну поведінку. Практична частина охоплює підготовку CSV-датасету, формування RFM-показників, кластеризацію методом k-means у середовищі R та оцінювання отриманих груп клієнтів. Результати подано у Power BI у вигляді dashboard з аналітичними показниками, візуалізаціями та висновками щодо особливостей кожного сегмента.

Ключові слова: сегментація клієнтів, RFM-аналіз, кластеризація, k-means, R, Power BI, бізнес-аналітика.

## ABSTRACT

**Nazarenko M. S.** Development of an Analytical System for Customer Segmentation Based on Statistical Learning Methods and Business Intelligence Tools. Specialty 122 "Computer Science". Vasyl Stus Donetsk National University, Vinnytsia, 2026.

The qualification work examines customer segmentation based on purchasing behavior data. The practical part includes CSV dataset preparation, calculation of RFM indicators, k-means clustering in R, and evaluation of the obtained customer groups. The results are presented in Power BI as a dashboard with analytical indicators, visualizations, and conclusions describing the characteristics of each segment.

Keywords: customer segmentation, RFM analysis, clustering, k-means, R, Power BI, business intelligence

# ЗМІСТ

# ВСТУП

**Актуальність теми дослідження.** Сучасний етап розвитку електронної комерції характеризується стрімким нарощуванням обсягів даних, які генеруються внаслідок взаємодії клієнтів із підприємствами. Кожна транзакція, кожна переглянута товарна позиція та залишений відгук формують ту первинну інформаційну основу, на якій вибудовується сучасний клієнт-орієнтований маркетинг. Здатність опрацьовувати, систематизувати та інтерпретувати такі дані стає одним із ключових чинників конкурентоспроможності e-commerce підприємств і визначає якість прийняття управлінських рішень.

Особливу роль у цьому процесі відіграє сегментація клієнтів – поділ клієнтської бази на однорідні групи з подібною купівельною поведінкою. Сегментація дозволяє переходити від уніфікованого маркетингу до персоналізованих стратегій взаємодії, оптимізувати рекламні витрати, утримувати найцінніших клієнтів та виявляти групи, що потребують реактивації. Водночас традиційні підходи до сегментації, що ґрунтуються на демографічних чи інтуїтивних критеріях, не розкривають усього потенціалу транзакційних даних та не забезпечують достатньої точності у визначенні поведінкових патернів.

Поглиблене вирішення цієї проблеми пов'язане із застосуванням методів статистичного навчання, передусім алгоритмів кластеризації без учителя. Метод K-means у поєднанні з RFM-аналізом (Recency, Frequency, Monetary) набув значного поширення в академічних дослідженнях та комерційній практиці завдяки прозорості, обчислювальній ефективності та зрозумілій бізнес-інтерпретації результатів. Інтеграція такої моделі з інструментами бізнес-аналітики, зокрема платформою Microsoft Power BI, дозволяє трансформувати математичні результати кластеризації у інтерактивні аналітичні рішення, доступні кінцевим користувачам без спеціальної технічної підготовки.

Незважаючи на широке висвітлення окремих елементів цього підходу у науковій літературі, побудова цілісних аналітичних систем, які поєднують

програмну реалізацію статистичного навчання, обробку реальних транзакційних даних і повноцінну візуалізацію у середовищі бізнес-аналітики, залишається актуальним завданням. Усе зазначене зумовлює актуальність розробки аналітичної системи сегментації клієнтів на основі методів статистичного навчання із застосуванням інструментів бізнес-аналітики.

**Мета дослідження** полягає у розробці аналітичної системи сегментації клієнтів e-commerce підприємств на основі методу K-means кластеризації та RFM-аналізу з реалізацією інтерактивного аналітичного дашборду у середовищі Power BI.

**Завдання дослідження** визначаються поставленою метою та полягають у такому:

- дослідити сутність і роль сегментації клієнтів у сучасному маркетингу, узагальнити підходи до аналізу клієнтських даних;
- розглянути теоретичні засади бізнес-аналітики, охарактеризувати її роль у формуванні data-driven підходу до прийняття управлінських рішень;
- проаналізувати методи статистичного навчання, що застосовуються у задачах сегментації, та обґрунтувати доцільність використання алгоритму K-means;
- розкрити концепцію RFM-аналізу та особливості його інтеграції з методами кластеризації;
- обґрунтувати вибір реального датасету транзакційних даних для практичної реалізації моделі;
- виконати попередню обробку, очищення та інтеграцію даних, сформувати показники Recency, Frequency, Monetary;
- побудувати кластерну модель методом K-means, визначити оптимальну кількість сегментів та оцінити якість кластеризації;

- реалізувати програмну частину аналітичної системи мовою R із використанням спеціалізованих бібліотек статистичного навчання;
- спроєктувати та розробити інтерактивний аналітичний дашборд у середовищі Power BI;
- сформулювати аналітичну інтерпретацію отриманих сегментів та практичні рекомендації щодо їхнього використання.

**Об'єкт дослідження** – процес сегментації клієнтів e-commerce підприємств на основі аналізу транзакційних даних.

**Предмет дослідження** – методи статистичного навчання та інструменти бізнес-аналітики, застосовані для побудови аналітичної системи сегментації клієнтів.

**Методи дослідження.** Для розв'язання поставлених завдань у роботі використано загальнонаукові методи аналізу, синтезу, порівняння та узагальнення – під час опрацювання теоретичних джерел та формулювання висновків; методи розвідувального аналізу даних (EDA), описової статистики, кореляційного аналізу – для дослідження структури транзакційного набору; методи попередньої обробки даних, нормалізації та логарифмічної трансформації – для підготовки показників до кластеризації; метод RFM-аналізу – для формування поведінкових показників клієнтів; алгоритм K-means кластеризації – для побудови сегментаційної моделі; методи Elbow, silhouette coefficient та gap statistic – для визначення оптимальної кількості кластерів; методи візуалізації даних та проєктування інтерактивних дашбордів засобами Power BI – для практичної реалізації аналітичної системи.

**Теоретичне та практичне значення** одержаних результатів. Теоретичне значення роботи полягає у систематизації сучасних підходів до сегментації клієнтів на основі методів статистичного навчання та обґрунтуванні доцільності інтеграції RFM-аналізу, K-means кластеризації й інструментів бізнес-аналітики у межах єдиної аналітичної системи. Практичне значення полягає у тому, що розроблена

аналітична система може бути використана e-commerce підприємствами для виявлення найцінніших груп клієнтів, оптимізації маркетингових кампаній, формування персоналізованих стратегій утримання споживачів та підвищення ефективності управлінських рішень. Програмна реалізація мовою R та дашборд у Power BI можуть бути адаптовані під реальні набори даних інших підприємств без істотних модифікацій.

**Апробація результатів дослідження**. Основні положення та результати дослідження висвітлено у науковій публікації: Назаренко М. С., Січко Т. В. Аналіз клієнтських даних із застосуванням інструментів бізнес-аналітики.

**Структура роботи.** Кваліфікаційна (бакалаврська) робота складається зі вступу, трьох розділів основної частини, висновків, списку використаних посилань та додатків. У першому розділі викладено теоретичні основи сегментації клієнтів, методів статистичного навчання та інструментів бізнес-аналітики. Другий розділ присвячено розробці моделі сегментації: обґрунтовано вибір датасету, описано процедуру попередньої обробки даних, формування RFM-показників та побудови кластерної моделі методом K-means. У третьому розділі представлено архітектуру аналітичної системи, її програмну реалізацію мовою R, проєктування та розробку інтерактивного дашборду у Power BI, а також аналітичну інтерпретацію результатів. Робота містить 60 сторінок основного тексту, 14 рисунків, 12 таблиць, 3 формули; перелік джерел налічує 43 найменування.

# РОЗДІЛ 1. ТЕОРЕТИЧНІ ОСНОВИ СЕГМЕНТАЦІЇ КЛІЄНТІВ ТА АНАЛІЗУ ДАНИХ У БІЗНЕС-АНАЛІТИЦІ

## 1.1. Сегментація клієнтів як інструмент клієнт-орієнтованого управління

Сегментація клієнтів становить методологічний підхід до аналізу клієнтської бази підприємства, що передбачає її поділ на відносно однорідні групи (сегменти) на основі схожості певних характеристик – демографічних, географічних, поведінкових або психографічних. Концепція бере витоки з робіт В. Сміта [1], який обґрунтував перехід від масового маркетингу до диференційованого, та Ф. Котлера, який сформулював класичну модель STP (Segmentation, Targeting, Positioning) як основу сучасного маркетингового планування [2]. Подальший розвиток підходу пов'язаний з появою концепцій управління відносинами з клієнтами та зростанням обчислювальних можливостей, що дозволили перейти від теоретичних побудов до практичної реалізації сегментаційних моделей на великих масивах транзакційних даних.

У межах клієнт-орієнтованого управління (Customer Relationship Management, CRM) сегментація виконує кілька взаємопов'язаних функцій. Передусім вона дає змогу розподіляти маркетингові комунікації відповідно до особливостей кожної групи, що підвищує релевантність повідомлень та коефіцієнт відгуку. Окрім цього, сегментація створює основу для розрахунку показника життєвої цінності клієнта (Customer Lifetime Value, CLV) у контексті окремих груп, що, у свою чергу, дозволяє оптимізувати розподіл маркетингового бюджету. Не менш важливою функцією є формування стратегій утримання клієнтів, реактивації неактивних споживачів та масштабування програм лояльності [3]. У такий спосіб сегментація перетворюється на наскрізний інструмент управлінського процесу – від оперативних рішень щодо окремих кампаній до стратегічного планування продуктової лінійки.

Актуальні підходи до сегментації прийнято поділяти на чотири основні типи: демографічна, географічна, поведінкова та психографічна сегментація [4]. Демографічна сегментація ґрунтується на статичних характеристиках споживачів (вік, стать, рівень доходу, освіта). Географічна враховує територіальну приналежність та особливості регіональних ринків. Психографічна охоплює стиль життя, цінності, особистісні характеристики споживачів. Поведінкова сегментація фіксує реальні дії клієнтів у взаємодії з підприємством – частоту купівель, обсяги витрат, переваги у виборі товарних категорій, давність останньої транзакції. В рамках електронної комерції саме поведінкова сегментація має найбільшу аналітичну цінність, оскільки вона спирається на об'єктивні дані транзакційної активності, які підприємство фіксує безпосередньо у власних інформаційних системах і які не вимагають додаткового збору соціологічної інформації.

Перехід від інтуїтивних до підходів на основі даних (data-driven) сегментації супроводжується істотним підвищенням точності виокремлення сегментів та обґрунтованості прийнятих рішень. У наукових публікаціях [5, 6] зазначається, що традиційні підходи, побудовані виключно на демографічних критеріях, у середовищі сучасного e-commerce поступово втрачають свою прогнозну здатність, оскільки покупці однієї вікової чи дохідної групи демонструють принципово відмінну поведінку. Натомість аналіз транзакційних даних – частоти покупок, обсягів витрат, давності останньої взаємодії – дозволяє виявляти стійкі поведінкові патерни, на основі яких формуються дієві маркетингові стратегії. Саме такий підхід, з опорою на поведінкові дані та статистичні методи, покладено в основу подальших розділів цього дослідження.

### 1.2. Аналіз клієнтських даних: типи, джерела, підходи

Поняття клієнтських даних охоплює сукупність структурованих та неструктурованих відомостей, які підприємство збирає, зберігає та обробляє стосовно своїх споживачів. Залежно від характеру та джерела походження

клієнтські дані прийнято класифікувати з урахуванням деяких вимірів. За змістом виокремлюють демографічні дані (стать, вік, місце проживання), транзакційні дані (історія купівель, обсяги, частота, продуктові категорії), поведінкові дані (відвідування сайту, переглянуті товари, реакція на рекламні комунікації), а також soft-data, що включають відгуки, оцінки, інтенсивність взаємодії в соціальних мережах [7]. За способом отримання дані поділяють на first-party (зібрані безпосередньо підприємством), second-party (отримані від партнерів) та third-party (придбані у спеціалізованих постачальників). Найбільшу аналітичну цінність становлять саме first-party транзакційні дані, оскільки вони характеризуються високою якістю та актуальністю.

Джерела клієнтських даних в e-commerce підприємствах є різноманітними. Основу формують внутрішні інформаційні системи: CRM-платформи, що містять профілі клієнтів та історію взаємодії, ERP-системи, які фіксують фінансові операції, та власне платформи електронної комерції, що генерують транзакційні логи. Додатковими джерелами слугують системи веб-аналітики (Google Analytics, Adobe Analytics), маркетингові платформи (інструменти email-маркетингу, рекламні кабінети), а також зовнішні API сервісів логістики, платіжних систем та соціальних мереж [8]. Об'єднання цих джерел у єдиний аналітичний контур потребує застосування ETL-процесів (Extract, Transform, Load) та централізованих сховищ даних (Data Warehouse), що становить технологічну основу подальшого аналізу.

Сучасна концепція опрацювання клієнтських даних реалізується через концепцію data-driven decision-making – підхід до прийняття управлінських рішень, який ставить емпіричні дані в центр аналітичного процесу та мінімізує вплив суб'єктивних суджень. Дослідники [9] виокремлюють кілька етапів розвитку data-driven підходу: описовий аналіз (descriptive analytics) дає відповідь на запитання «що відбулося?», діагностичний (diagnostic) – «чому це сталося?», прогностичний (predictive) – «що може статися?», та рекомендаційний (prescriptive) – «що робити?». Сегментація клієнтів традиційно перебуває на межі описового та

діагностичного рівнів, однак у поєднанні з методами статистичного навчання вона набуває рис прогностичного аналізу, оскільки виокремлені сегменти стають основою для передбачення майбутньої поведінки клієнтів.

Особливою рисою сучасного аналізу клієнтських даних є його циклічна природа. Жоден сегментаційний результат не є остаточним – клієнтська база перебуває в постійному русі, нові клієнти приходять, існуючі змінюють поведінку, частина переходить у статус неактивних. Це зумовлює необхідність регулярного оновлення сегментаційної моделі та інтеграції аналітики в операційні бізнес-процеси. На технологічному рівні така інтеграція забезпечується інструментами бізнес-аналітики, які поєднують у собі функції обробки даних, статистичного моделювання та інтерактивної візуалізації.

### 1.3. Бізнес-аналітика як технологічна основа сегментації

Бізнес-аналітика (Business Intelligence, BI) – це сукупність методологій, процесів, архітектур та технологій, що перетворюють великі масиви даних на корисну для прийняття управлінських рішень інформацію [10]. Концептуально BI охоплює весь життєвий цикл роботи з даними: від їхнього збору з джерел операційного обліку до представлення кінцевим користувачам у вигляді інтерактивних звітів і дашбордів. Сучасне розуміння бізнес-аналітики передбачає не лише ретроспективний аналіз історичних даних, а й елементи прогнозування, моделювання сценаріїв та інтеграції з методами машинного навчання, що робить її ключовою технологічною основою для побудови аналітичних систем сегментації клієнтів.

Архітектура типової BI-системи включає кілька функціональних шарів. На рівні джерел даних розташовуються операційні системи (ERP, CRM, e-commerce платформи), бази транзакцій, файли різноманітних форматів та зовнішні API. Шар інтеграції реалізує ETL-процеси: вилучення даних із джерел (Extract), їхнє очищення, трансформацію, узгодження форматів (Transform) та завантаження до

центрального сховища (Load). Шар зберігання представлений сховищем даних (Data Warehouse) або, у сучасних архітектурах, озером даних (Data Lake), де інформація зберігається в нормалізованому або напівструктурованому вигляді. На рівні аналітичної обробки виконуються запити, статистичні розрахунки та моделювання. Завершальним шаром є презентація – інтерактивні дашборди, звіти, аналітичні панелі, призначені для кінцевих користувачів [11].

Сучасний ринок BI-платформ представлений широким спектром рішень, серед яких найбільш поширеними є Microsoft Power BI, Tableau, Qlik Sense та Google Looker. Усі вони реалізують спільну функціональну логіку – підключення до джерел, моделювання даних, побудова візуалізацій, – однак відрізняються архітектурними особливостями, ціновою політикою та інтеграційними можливостями. Узагальнене порівняння цих платформ за ключовими характеристиками наведено в табл. 1.1.

Таблиця 1.1 – Порівняльна характеристика основних BI-платформ

| **Критерій** | **Power BI** | **Tableau** | **Qlik Sense** | **Looker** |
|---|---|---|---|---|
| Розробник | Microsoft | Salesforce | Qlik | Google |
| Модель ліцензування | Subscription / Free Desktop | Subscription | Subscription | Subscription |
| Мова формул | DAX | VizQL | Qlik Script | LookML |
| Інтеграція з R / Python | Висока | Висока | Середня | Обмежена |
| Доступність для малого бізнесу | Висока | Середня | Середня | Низька |
| Спільнота користувачів | Дуже велика | Велика | Середня | Зростаюча |

З урахуванням наведених у таблиці характеристик у межах цього дослідження як інструмент реалізації аналітичного дашборду обрано Microsoft Power BI. На користь такого вибору свідчать кілька чинників: безкоштовна доступність Power BI Desktop, можливість підключення до різноманітних джерел даних, зокрема файлів формату CSV, потужна модель формул DAX, активна спільнота користувачів, а

також орієнтованість платформи на сектор малого та середнього бізнесу, який може становити цільову аудиторію розроблюваної системи. Детальний розгляд функціональних можливостей Power BI наведено у підрозділі 1.8.

### 1.4. Методи статистичного навчання в задачах сегментації

Статистичне навчання (statistical learning) – це галузь прикладної статистики та машинного навчання, що об'єднує методи виявлення закономірностей у даних та побудови моделей для прогнозу або опису. Основу класифікації методів статистичного навчання становить наявність або відсутність маркованих даних, на основі чого виокремлюють два великі класи: навчання з учителем або кероване навчання (supervised learning) та навчання без учителя (unsupervised learning) [12]. Третій клас – навчання з підкріпленням дій (reinforcement learning) – лежить поза межами цього дослідження, оскільки не застосовується в типових задачах сегментації.

Навчання з учителем передбачає наявність набору даних, у якому для кожного спостереження відомий правильний результат (мітка класу або значення цільової змінної). Завданням алгоритму є побудова функціональної залежності між вхідними ознаками та цільовою змінною, яка може бути використана для прогнозування на нових даних. До методів керованого навчання належать регресія, дерева рішень, опорно-векторні машини, нейронні мережі та інші. У контексті аналізу клієнтських даних такі методи застосовують для задач, у яких відомий цільовий клас: прогнозування відтоку клієнтів (churn prediction), оцінювання ймовірності купівлі, передбачення значення CLV. Однак для задачі сегментації клієнтів такий тип навчання не є природним вибором, оскільки розмітка клієнтів за «правильними» сегментами відсутня – самі сегменти потрібно виявити в процесі аналізу.

Навчання без учителя оперує немічеными даними та має на меті виявлення прихованої структури в наборі. Основними типами задач є кластеризація

(clustering), зниження розмірності (dimensionality reduction), виявлення асоціативних правил та аномалій. Саме методи кластеризації становлять методологічну основу сегментації клієнтів побудованих на даних: на вхід подається матриця ознак клієнтів (наприклад, RFM-показників), а на виході отримується розбиття на групи без попереднього знання про їхню кількість чи структуру [13]. Цей підхід відповідає природі задачі сегментації, у якій аналітик прагне виявити природні групи всередині клієнтської бази, а не класифікувати клієнтів за заздалегідь визначеними категоріями.

Доцільність застосування саме методів некерованого навчання в задачі сегментації клієнтів обґрунтована кількома міркуваннями. По-перше, відсутні апріорні мітки сегментів, тому будь-яка спроба використати supervised-методи вимагала б штучного формування міток, що знизило б об'єктивність аналізу. По-друге, кластерні методи дозволяють виявити нетипові структури в даних, які неочевидні з простих описових статистик. По-третє, отримані сегменти інтерпретуються в термінах характеристик кожної групи, що зручно для подальшого формування маркетингових стратегій. Серед усього розмаїття алгоритмів кластеризації для задачі сегментації клієнтів на основі RFM-показників найбільш доцільним є алгоритм K-means, обґрунтування вибору якого наводиться в наступному підрозділі.

### 1.5. Кластеризація як основний метод сегментації клієнтів

Кластеризація – це задача поділу множини об'єктів на групи (кластери) таким чином, щоб об'єкти всередині одного кластера були максимально схожі між собою, а об'єкти різних кластерів – максимально відмінні. Формально, маючи набір з n спостережень, кожне з яких описується вектором ознак у p-вимірному просторі, алгоритм кластеризації будує функцію відображення цих спостережень у k непересічних підмножин на основі обраної метрики подібності або відстані [14]. У задачі сегментації клієнтів об'єктами кластеризації виступають окремі клієнти,

ознаками – їхні поведінкові характеристики (зокрема, RFM-показники), а кластерами – цільові сегменти клієнтської бази.

Сучасні алгоритми кластеризації прийнято класифікувати за принципом їхньої роботи на кілька основних груп. Методи розподілу (зокрема K-means, K-medoids) розбивають набір на наперед задану кількість кластерів через ітеративну оптимізацію цільової функції. Ієрархічні методи (агломеративна та divisive кластеризація) будують ієрархію вкладених кластерів, що візуалізується у вигляді дендрограми. Density-based методи (DBSCAN, OPTICS) визначають кластери як області підвищеної щільності точок у просторі ознак. Модельні методи (gaussian mixture models) припускають, що дані породжені сумішшю ймовірнісних розподілів та оцінюють їхні параметри [15]. Кожен клас методів має переваги та обмеження залежно від характеру задачі.

Узагальнене порівняння найпоширеніших алгоритмів кластеризації у контексті задачі customer segmentation наведено в табл. 1.2.

Таблиця 1.2 – Порівняння алгоритмів кластеризації для задач сегментації клієнтів

| **Алгоритм** | **Тип** | **Потребує задання K** | **Стійкість до викидів** | **Доцільність для RFM** |
|---|---|---|---|---|
| K-means | Розбиття | Так | Низька | Висока |
| K-medoids (PAM) | Розбиття | Так | Висока | Середня |
| Hierarchical | Ієрархічний | Ні | Середня | Середня |
| DBSCAN | Density-based | Ні | Висока | Низька |
| GMM | Модельний | Так | Середня | Середня |

Аналіз наведених характеристик дозволяє обґрунтувати вибір алгоритму K-means для задачі сегментації клієнтів за RFM-показниками. Перевагами K-means у цьому контексті є висока обчислювальна ефективність (важлива при обробці великих транзакційних масивів), визначена поведінка при фіксованому початковому стані, простота інтерпретації результатів через центроїди кластерів,

широка підтримка у статистичних пакетах і платформах. Обмеження K-means – необхідність попереднього задання кількості кластерів та чутливість до викидів – нівелюються відповідними процедурами підготовки даних (логарифмічна трансформація, масштабування) та застосуванням спеціальних методів визначення оптимального K, що детально розглядаються в наступному підрозділі. Порівняльне дослідження кластерних алгоритмів для customer segmentation на даних британського ритейлу [16] підтверджує перевагу K-means як базової моделі.

### 1.6. Алгоритм K-means: суть, особливості, методи вибору кількості кластерів

Алгоритм K-means запропонований С. Ллойдом у середині минулого століття та залишається одним із найпопулярніших методів кластеризації завдяки своїй простоті та ефективності [17]. Методологічно алгоритм базується на принципі мінімізації внутрішньокластерної дисперсії: набір з n спостережень розбивається на k кластерів таким чином, щоб сумарна квадратична відстань від кожної точки до центроїда її кластера була мінімальною. Цільова функція алгоритму записується у вигляді (1.1):

$$\mathrm{WCSS} = \Sigma\,\Sigma\,||x_j - \mu_i||^2 \tag{1.1}$$

де WCSS – within-cluster sum of squares (сумарна внутрішньокластерна дисперсія);

$x_j$ – спостереження, що належить кластеру i;

$\mu_i$ – центроїд кластера i.

Робота алгоритму складається з кількох послідовних кроків. На початковому етапі обирається кількість кластерів k та ініціалізуються k центроїдів (зазвичай випадковим чином або з використанням евристики k-means++). На наступному кроці кожне спостереження відносять до того кластера, центроїд якого є найближчим за обраною метрикою (зазвичай – евклідовою відстанню). Після цього

центроїди перераховуються як середні значення ознак усіх спостережень, що належать до відповідного кластера. Кроки призначення та перерахунку повторюються ітеративно до моменту, коли центроїди перестають змінюватися (або зміни стають меншими за обраний поріг). Такий ітераційний процес гарантує монотонне зменшення цільової функції, однак не гарантує знаходження глобального мінімуму – алгоритм може зупинитися в локальному оптимумі залежно від початкової ініціалізації центроїдів.

Однією з ключових практичних проблем застосування K-means є визначення оптимальної кількості кластерів k, яке не задається алгоритмом автоматично та потребує окремої процедури обґрунтування. У відповідній літературі [18; 19] описано кілька основних методів вибору K, що використовуються взаємодоповнювально.

Метод ліктя (Elbow method) базується на побудові графіка залежності WCSS від кількості кластерів k та виявленні точки «перегину», після якої подальше збільшення k не призводить до істотного зниження дисперсії. Цей метод інтуїтивно зрозумілий, однак точка перегину не завжди є очевидною, особливо для реальних транзакційних даних із непомітною кластерною структурою.

Коефіцієнт силуету (Silhouette coefficient) для кожного спостереження вимірює, наскільки воно подібне до інших точок свого кластера порівняно з найближчим сусіднім кластером, та обчислюється за формулою (1.2):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{1.2}$$

де a(i) – середня відстань точки i до інших точок свого кластера;

b(i) – середня відстань точки i до точок найближчого сусіднього кластера.

Значення s(i) варіюється від −1 до 1: позитивні значення вказують на коректне віднесення точки до кластера, від'ємні – на ймовірну помилку призначення. Середнє

значення коефіцієнта силуету по всьому набору використовується як зведена метрика якості кластеризації для різних значень k.

Gap statistic (статистика розриву), запропонована Р. Тібшірані та співавторами, порівнює WCSS отриманої кластеризації з очікуваним WCSS за нульової гіпотези про відсутність кластерної структури (тобто за рівномірного розподілу даних). Оптимальне K відповідає найбільшому розриву між фактичним та очікуваним значеннями WCSS [19].

У практиці побудови сегментаційних моделей зазвичай застосовують одночасно кілька з наведених методів, що дозволяє підвищити обґрунтованість вибору K. У контексті цього дослідження для визначення оптимальної кількості кластерів буде використано комбінацію методу ліктя та коефіцієнта силуету, що детально описано в підрозділі 2.8.

Окрім вибору K, ефективність застосування K-means істотно залежить від попередньої підготовки даних. Зокрема, оскільки алгоритм використовує евклідову відстань, ознаки з більшим діапазоном значень будуть домінувати в розрахунку відстаней. Це робить необхідним масштабування ознак, що зазвичай реалізується через z-score стандартизацію. Додатково для RFM-показників, які зазвичай мають сильну правосторонній нахил розподілу, доцільним є застосування логарифмічної трансформації, що наближає розподіли до симетричних та підвищує якість кластеризації [20].

### 1.7. RFM-аналіз: концепція та інтеграція з методами кластеризації

RFM-аналіз – це методологічний підхід до сегментації клієнтів, що ґрунтується на трьох поведінкових показниках: Recency (давність останньої транзакції), Frequency (частота транзакцій) та Monetary (сумарний грошовий обсяг покупок). Підхід запропоновано А. Хьюзом у 1994 році в рамках теорії database marketing та з того часу набув статусу класичного інструменту customer analytics завдяки інтуїтивності, доступності та доведеній на практиці ефективності [21].

Концептуальне обґрунтування RFM-аналізу базується на спостереженні: клієнт, що нещодавно зробив покупку, купує часто та витрачає значні суми, є найціннішим для підприємства; відповідно, клієнти, які давно не купували та витрачали мало, потребують або реактивації, або виключення з активних маркетингових кампаній.

Кожен із трьох показників RFM має конкретний обчислювальний зміст у контексті транзакційних даних. Recency визначається як кількість днів (або іншої одиниці часу) між референсною датою аналізу та датою останньої транзакції клієнта. Frequency – кількість унікальних транзакцій клієнта за обраний період аналізу. Monetary – сумарна вартість усіх покупок клієнта за той самий період.

Класична реалізація RFM-аналізу передбачає наступну схему: кожен із трьох показників розбивається на квартилі та клієнтові присвоюється бал від 1 до 5 за кожною з осей. Це дає 125 можливих RFM-комбінацій, які перетворюються у бізнес-сегменти на кшталт «Champions», «Loyal Customers», «At Risk», «Hibernating». Незважаючи на свою практичну цінність, такий підхід має істотні обмеження: межі сегментів встановлюються штучно через квартильні границі, а реальна структура клієнтської бази може не відповідати рівномірному розподілу за квартилями [22]. Окрім цього, така модель не враховує відносної ваги показників та потенційних нелінійних залежностей між ними.

Подоланням наведених обмежень є інтеграція RFM-аналізу з методами кластеризації, зокрема з алгоритмом K-means. У такому підході RFM-показники використовуються як ознаки клієнтів у просторі кластеризації, а сегменти формуються не штучними квантильними границями, а природною структурою даних. Це дозволяє виявляти сегменти змінного розміру, які реально присутні в клієнтській базі, та отримувати інтерпретовні центроїди кластерів у термінах середніх значень R, F, M по кожній групі [23]. Останніми роками такий гібридний підхід (RFM + K-means) став стандартом для академічних досліджень та комерційних рішень у сфері customer segmentation [24].

Типові сегменти, що отримуються в результаті RFM+K-means кластеризації, мають усталені бізнес-назви, які відображають їхні характеристики. До найпоширеніших належать Champions – клієнти з низьким Recency, високим Frequency та високим Monetary; Loyal Customers – стабільно активні клієнти з помірно високими показниками за всіма осями; Potential Loyalists – нещодавні клієнти з помірними Frequency та Monetary, що мають потенціал зростання; At Risk – клієнти з високим Recency, але історично високими Frequency та Monetary, які перестали купувати; Hibernating або Lost – клієнти з тривалою відсутністю транзакцій та низькими історичними показниками [25]. Така типологія є орієнтовною – конкретні сегменти, їхня кількість та назви визначаються специфікою клієнтської бази та результатами кластеризації.

Інтеграція RFM-аналізу та K-means кластеризації становить методологічну основу системи сегментації, що розробляється в цій кваліфікаційній роботі. Така інтеграція забезпечує поєднання інтерпретованості класичного RFM-підходу зі статистичною обґрунтованістю кластерного аналізу та створює міцну основу для майбутньої візуалізації результатів у середовищі Power BI.

### 1.8. Power BI як інструмент реалізації аналітичного дашборду

Microsoft Power BI – це інтегрована платформа бізнес-аналітики, що поєднує функції підключення до різноманітних джерел даних, їхньої трансформації, моделювання та інтерактивної візуалізації. Платформа реалізована у кількох варіантах: Power BI Desktop – безкоштовна версія для розробки моделей та дашбордів; Power BI Service – хмарне середовище для публікації, спільного використання та автоматичного оновлення звітів; Power BI Mobile – мобільні застосунки для перегляду дашбордів [26]. Така архітектура гарантує повний життєвий цикл аналітичного рішення – від розробки до доставки результатів кінцевому користувачеві.

Функціональна архітектура Power BI базується на трьох взаємопов'язаних компонентах. Power Query – інструмент ETL, що дозволяє підключатися до сотень типів джерел (бази даних, файли, веб-сервіси, API), виконувати трансформації даних та формувати фінальну модель. Усі трансформації записуються у форматі мови M, що забезпечує повторюваність та автоматизацію оновлення. DAX (Data Analysis Expressions) – мова формул для створення обчислюваних стовпців, мір та таблиць у моделі даних. DAX забезпечує функціональність, аналогічну формулам Excel, однак працює з повноцінною реляційною моделлю та підтримує складні аналітичні розрахунки (time intelligence, контекстне фільтрування, ієрархічна агрегація). Power View – шар візуалізації, що надає широкий набір типів графіків, карток KPI, таблиць, географічних карт та можливість створення власних візуальних компонентів [29].

Ключовою перевагою Power BI для задач customer analytics є можливість створення інтерактивних дашбордів, у яких візуалізації пов'язані між собою через механізм cross-filtering. Це означає, що вибір значення на одному графіку автоматично фільтрує дані на всіх інших візуалізаціях дашборду. Така інтерактивність дає змогу аналітику динамічно досліджувати дані в різних розрізах – за сегментами клієнтів, географічними регіонами, часовими періодами – без необхідності формування статичних звітів для кожного запиту. Цю властивість Power BI було продемонстровано у науковій публікації [28], де на прикладі транзакцій інтернет-магазину показано, як вибір сегмента в круговій діаграмі автоматично перебудовує всі інші елементи дашборду, забезпечуючи деталізований аналіз окремих груп споживачів.

Особливе значення для аналітичної системи, що розробляється у цій роботі, має спосіб обміну даними між середовищем статистичних обчислень та платформою візуалізації. У межах проєкту обрано підхід, за якого статистична обробка та кластеризація повністю виконуються у середовищі R (RStudio), а результати сегментації експортуються у проміжний CSV-файл, який імпортується

у Power BI як джерело даних. Такий підхід забезпечує чіткий розподіл відповідальності між компонентами: R відповідає за обчислювально складні етапи статистичного навчання, а Power BI – за інтерактивну візуалізацію та представлення результатів кінцевому користувачеві. Перевагою рішення є незалежність етапів аналізу й візуалізації, простота відтворення та можливість оновлення сегментації шляхом повторного запуску R-скрипту.

### 1.9. Огляд сучасних практик customer analytics в e-commerce

Сучасний стан customer analytics характеризується кількома взаємопов'язаними тенденціями, що визначають вектор розвитку аналітичних систем сегментації клієнтів. Першою з них є перехід від ретроспективного аналізу до аналітики в реальному часі (real-time analytics). Якщо традиційні BI-системи оперують даними, оновлюваними з періодичністю «день – тиждень – місяць», то сучасні рішення прагнуть забезпечити аналіз поведінки клієнтів безпосередньо в момент її здійснення. Це досягається через streaming-архітектури (Apache Kafka, AWS Kinesis), системи in-memory обчислень та інтеграцію аналітичних результатів безпосередньо в операційні системи – рекомендаційні сервіси, маркетингові платформи, чат-боти [29].

Другою важливою тенденцією є predictive segmentation – поєднання класичної сегментації з прогнозованими моделями. Замість виокремлення сегментів виключно на основі історичних даних, predictive-підходи прогнозують, до якого сегмента клієнт належатиме в майбутньому, або яка ймовірність переходу між сегментами. Це дозволяє формувати випереджальні маркетингові стратегії – реактивувати потенційно неактивних клієнтів до того, як вони фактично припинять купівлі, або стимулювати лояльних до переходу в сегмент Champions [30]. Такі моделі зазвичай реалізуються через комбінацію кластеризації та класифікаційних алгоритмів (логістична регресія, random forest, gradient boosting), однак потребують

значно більшої кількості даних та обчислювальних ресурсів порівняно з класичним RFM-підходом. Сучасні порівняльні дослідження алгоритмів кластеризації для сегментації клієнтів [31] підтверджують стабільність методу K-means як базової моделі в e-commerce задачах.

Третьою тенденцією є інтеграція штучного інтелекту в BI-платформи (AI-augmented BI). Сучасні версії Power BI, Tableau та Qlik містять вбудовані функції автоматичного виявлення інсайтів, генерації пояснень для аномалій, обробки запитів природною мовою (natural language Q&A). Це знижує бар'єр входу для бізнес-користувачів та дозволяє отримувати аналітичні висновки без глибоких технічних знань. Водночас базові методи статистичного аналізу (зокрема, RFM та K-means) залишаються прозорими, інтерпретованими та контрольованими – що становить їхню перевагу у задачах, де важливо обґрунтувати кожен крок аналізу [32].

Кейси провідних e-commerce платформ демонструють практичну реалізацію наведених тенденцій. Amazon використовує сегментацію клієнтів як основу системи рекомендацій, об'єднуючи поведінкові, демографічні та контекстні дані. Бразильська платформа Mercado Libre інтегрує сегментаційні моделі в маркетингові комунікації по всій Латинській Америці[33]. Український ринок електронної комерції, представлений такими гравцями, як Rozetka та Prom.ua, поступово переходить до даних керованого підходу, однак значна частина малого та середнього бізнесу досі використовує інтуїтивні методи сегментації або не сегментує клієнтську базу взагалі. Це зумовлює актуальність розробки доступних аналітичних рішень на основі відкритих інструментів (R та Power BI Desktop), що становить практичну спрямованість цього дослідження.

Окремої уваги заслуговує проблема якості даних, яка залишається ключовим обмежувальним чинником у customer analytics. Дослідження показують [34], що до 60–80% часу аналітика витрачається саме на підготовку та очищення даних, а не на

побудову моделей. Це підкреслює важливість методичного підходу до етапу попередньої обробки даних, який детально розглядається у Розділі 2.

**Висновки до розділу 1**

У першому розділі систематизовано теоретичні основи сегментації клієнтів та аналізу даних у бізнес-аналітиці. Обґрунтовано перевагу поведінкової сегментації над демографічною в умовах сучасного e-commerce та охарактеризовано концепцію data-driven decision-making як методологічну основу сучасних аналітичних систем. За результатами порівняльного аналізу провідних BI-платформ як інструмент реалізації обрано Microsoft Power BI.

Структуризовано методи статистичного навчання та обґрунтовано доцільність застосування алгоритмів кластеризації без учителя для задач сегментації. На основі порівняння K-means, K-medoids, ієрархічних, density-based та модельних методів обрано K-means як найбільш доцільний для сегментації за RFM-показниками; розглянуто методи визначення оптимальної кількості кластерів (Elbow, silhouette coefficient) та особливості попередньої підготовки даних. Розкрито концепцію RFM-аналізу і сучасний гібридний підхід RFM + K-means, який обрано як базовий метод системи.

Сформовані у цьому розділі теоретичні засади становлять основу для розробки моделі сегментації у Розділі 2 та реалізації аналітичної системи у Розділі 3.

# РОЗДІЛ 2. РОЗРОБКА МОДЕЛІ СЕГМЕНТАЦІЇ КЛІЄНТІВ НА ОСНОВІ МЕТОДІВ КЛАСТЕРИЗАЦІЇ

## 2.1. Обґрунтування вибору датасету для дослідження

Якість сегментаційної моделі безпосередньо залежить від характеру вхідних даних. Для практичної реалізації аналітичної системи сегментації клієнтів вибір датасету здійснювався на основі сформованих критеріїв: реальний характер транзакцій (без штучного генерування), наявність полів, придатних для формування RFM-показників, достатній обсяг для статистично обґрунтованої кластеризації, відкритий доступ для академічного використання, документована методологія збору, придатність до візуалізації у бізнес-аналітичних інструментах.

Серед розглянутих альтернатив порівняно три кандидати: Online Retail II Data Set (UCI Machine Learning Repository), Brazilian E-Commerce Public Dataset by Olist (Kaggle) та H&M Personalized Fashion Recommendations Dataset (Kaggle). Dataset Online Retail II має тривалу історію академічного застосування, однак охоплює період 2009–2011 років і станом на 2026 рік є застарілим для демонстрації сучасних аналітичних підходів. Dataset H&M містить дані за 2018–2020 роки та значний обсяг (понад 31 млн транзакцій), однак його структура потребує семплювання та значних обчислювальних ресурсів. Dataset Brazilian E-Commerce by Olist охоплює період вересень 2016 – жовтень 2018, містить структуровані транзакційні дані середнього масштабу, що оптимально підходить для академічного дослідження.

Підсумкове порівняння наведено в табл. 2.1.

Таблиця 2.1 – Порівняння кандидатних датасетів для сегментації клієнтів

| **Критерій** | **Online Retail II** | **Olist Brazilian E-Commerce** | **H&M Fashion** |
|---|---|---|---|
| Період охоплення | 2009–2011 | 2016–2018 | 2018–2020 |
| Обсяг (замовлень) | ~1 067 000 | ~99 000 | ~31 млн |
| Готовність до RFM | Висока | Висока | Середня (потребує семплювання) |
| Складність структури | 1 таблиця | 9 таблиць (join) | Багатотабличний |

| | | | |
|---|---|---|---|
| Реальність даних | Реальний UK-ритейлер | Реальний маркетплейс | Реальний бренд |
| Сучасність | Низька | Середня | Висока |

З урахуванням наведених характеристик для практичної реалізації аналітичної системи обрано Brazilian E-Commerce Public Dataset by Olist [35]. Цей датасет забезпечує оптимальне поєднання реальності даних, керованого обсягу, придатної для RFM структури та сучасності збору. Додатковим аргументом виступає висока цитованість набору у наукових публікаціях з customer analytics, що дозволяє порівнювати результати з іншими дослідженнями.

Olist – провідна бразильська платформа маркетплейс-моделі, що поєднує продавців із бразильських штатів із покупцями по всій країні. Бразилія входить до десятки найбільших ринків електронної комерції світу [36], що робить цей dataset репрезентативним. Dataset публічно доступний на платформі Kaggle та містить понад 99 тисяч замовлень за період вересень 2016 – жовтень 2018, що дає можливість сформувати повноцінні поведінкові показники клієнтів та провести статистично обґрунтовану кластеризацію.

## 2.2. Характеристика структури та змісту даних Brazilian E-Commerce by Olist

Brazilian E-Commerce Dataset by Olist – це набір даних, що складається з дев'яти CSV-файлів, які пов'язані між собою через зовнішні ключі. Така структура відображає типову схему транзакційної бази e-commerce маркетплейсу та потребує етапу інтеграції перед формуванням RFM-показників. Загальна характеристика таблиць набору даних подана у табл. 2.2.

Таблиця 2.2 – Опис таблиць dataset Brazilian E-Commerce by Olist

| **Таблиця** | **Записів** | **Ключові поля** | **Призначення** |
|---|---|---|---|
| olist_orders_dataset | 99 441 | order_id, customer_id, order_status, order_purchase_timestamp | Реєстр замовлень із часовими мітками |
| olist_customers_dataset | 99 441 | customer_id, customer_unique_id, customer_state | Профілі клієнтів |
| olist_order_items_dataset | 112 650 | order_id, product_id, price, freight_value | Позиції замовлень |
| olist_order_payments_dataset | 103 886 | order_id, payment_type, payment_value | Платіжні операції |
| olist_order_reviews_dataset | 99 224 | order_id, review_score | Оцінки клієнтів |
| olist_products_dataset | 32 951 | product_id, product_category_name | Каталог товарів |
| olist_sellers_dataset | 3 095 | seller_id, seller_state | Реєстр продавців |
| olist_geolocation_dataset | 1 000 163 | zip_code_prefix, lat, lng | Географічні координати |
| product_category_name_translation | 71 | category, category_en | Переклад назв категорій |

Структура зв'язків між таблицями характеризується типовою реляційною моделлю «замовлення – клієнт – позиції – платіж – відгук». Центральною таблицею є olist_orders_dataset, яка через поле customer_id пов'язана з olist_customers_dataset, а через order_id – з таблицями позицій, платежів та відгуків. Така архітектура дозволяє послідовно узагальнювати дані від рівня окремої позиції до рівня клієнта.

Особливо важливою для коректної сегментації є відмінність між полями customer_id та customer_unique_id у таблиці клієнтів. Поле customer_id є унікальним ідентифікатором у межах одного замовлення (формується для кожної транзакції окремо), тоді як customer_unique_id ідентифікує фізичного клієнта незалежно від кількості його замовлень. Загальна кількість записів у таблиці клієнтів становить 99 441, тоді як кількість унікальних клієнтів – 96 096. Це означає, що приблизно 3,5 %

клієнтів здійснили більше одного замовлення, що є нормальним показником для бразильського ринку маркетплейс-платформ.

Період охоплення датасету становить від 04 вересня 2016 року до 17 жовтня 2018 року, що відповідає приблизно 25 місяцям ринкової активності. Така тривалість достатня для коректного обчислення показника Recency без артефактів короткого вікна спостереження. Географічно набір даних покриває всі 27 штатів Бразилії з домінуванням південно-східного регіону: на штати São Paulo, Rio de Janeiro та Minas Gerais припадає сумарно понад 66 тис. записів, що становить 67 % клієнтської бази.

### 2.3. Попередня обробка та очищення даних

Якість будь-якої сегментаційної моделі критично залежить від якості вхідних даних, тому етап попередньої обробки є одним із найвідповідальніших у дослідженні. Згідно з оцінками [34], близько 60–80 % часу аналітика витрачається саме на цей етап. Для Olist було розроблено послідовний процес очищення, що передбачає кілька логічних кроків.

Перший крок – фільтрація замовлень за статусом доставки. Поле order_status містить вісім можливих значень: delivered (96 478 записів), shipped (1 107), canceled (625), unavailable (609), invoiced (314), processing (301), created (5), approved (2). Для коректного RFM-аналізу необхідно враховувати лише ті замовлення, які фактично завершилися доставкою товару клієнту, оскільки лише вони відповідають реальній купівельній активності та фінансовому результату. У результаті фільтрації за статусом delivered у вибірці залишилось 96 478 замовлень, що становить 97,02 % від загальної кількості.

Другий крок – обробка пропущених значень у часових мітках. У таблиці замовлень наявні чотири часові поля: order_purchase_timestamp (без пропусків), order_approved_at (160 пропусків), order_delivered_carrier_date (1 783 пропуски), order_delivered_customer_date (2 965 пропусків). Для RFM-аналізу критичним є саме

поле order_purchase_timestamp (дата покупки), яке не містить пропусків. Додатково проведено фільтрацію за наявністю order_delivered_customer_date для гарантії доставки, у результаті чого вибірка скоротилась до 96 470 замовлень.

Третій крок – перетворення типів даних. Усі часові поля під час імпорту з CSV є рядковими, тому їх необхідно конвертувати у формат datetime. Числові поля price, freight_value, payment_value мають коректний тип float64 та не потребують перетворення. Поле customer_state містить дволітерні коди штатів Бразилії та зберігається як object (string), що відповідає очікуваному формату для категоріальної ознаки.

Четвертий крок – перевірка наявності аномальних значень у фінансових полях. Аналіз показав, що мінімальне значення price складає R$ 0,85, максимальне – R$ 6 735,00. Середня вартість позиції замовлення – R$ 120,65, медіанна – R$ 74,99, що свідчить про правосторонню асиметрію розподілу цін. Жодних від'ємних значень або очевидних помилок введення не виявлено, що підтверджує високу якість первинного збору даних.

П'ятий крок – перевірка цілісності зв'язків між таблицями. Аналіз показав, що 100 % замовлень у статусі delivered мають відповідні записи у таблиці order_items, та аналогічно – у таблиці платежів. Це означає, що після фільтрації за статусом доставки додаткових втрат записів при join-операціях не очікується.

Підсумкова характеристика очищеного набору транзакцій наведена у табл. 2.3.

Таблиця 2.3 – Результати попередньої обробки даних

| **Етап обробки** | **Кількість замовлень** | **Зміна, %** |
|---|---|---|
| Початковий обсяг | 99 441 | – |
| Після фільтрації за статусом delivered | 96 478 | −2,98 % |
| Після видалення пропусків order_delivered_customer_date | 96 470 | −0,01 % |
| Після join з таблицею клієнтів | 96 470 | 0 % |
| Після join з таблицею позицій | 96 470 | 0 % |

| | | |
|---|---|---|
| Унікальних клієнтів (за customer_unique_id) | 93 350 | – |

### 2.4. Інтеграція даних: формування єдиного транзакційного набору

Реляційна структура Olist потребує послідовного об'єднання таблиць для формування єдиного транзакційного набору, придатного для RFM-аналізу. Інтеграція виконується через ланцюг join-операцій, при якому кожен наступний крок розширює доступну інформацію за рахунок приєднання нових атрибутів.

Перший рівень інтеграції – об'єднання таблиць orders та customers. Зв'язок здійснюється через поле customer_id, наявне в обох таблицях. У результаті формується розширена таблиця, у якій кожне замовлення асоційоване з конкретним клієнтом, його штатом та містом проживання. Цей крок дозволяє надалі групувати замовлення за фізичним клієнтом через поле customer_unique_id.

Другий рівень – агрегація даних на рівні позицій замовлення. Таблиця olist_order_items_dataset містить по одному запису для кожної товарної позиції в замовленні, причому одне замовлення може містити кілька позицій. Перед join з основною таблицею виконується агрегація по полю order_id з обчисленням сум: суми цін товарів (sum of price), сум доставки (sum of freight_value), кількості позицій (count). Це дозволяє звести інформацію до рівня замовлення та сформувати агреговану величину order_total = sum(price) + sum(freight_value), яка відображає повний грошовий обсяг транзакції.

Третій рівень – join розширеної таблиці orders+customers із узагальненою таблицею позицій. Зв'язок здійснюється через order_id. Результатом є транзакційна таблиця, яка містить для кожного замовлення: ідентифікатор клієнта (customer_unique_id), географічну приналежність (customer_state), часову мітку (order_purchase_timestamp), грошовий обсяг (order_total) та кількість позицій (n_items).

Загальна структура процесу інтеграції відображена у поданій нижче схемі (рис. 2.1). Кожен прямокутник відповідає окремій таблиці датасету, стрілки – операціям join, ромби – проміжним агрегаціям. Підсумковий результат – єдина таблиця транзакцій – слугує основою для формування RFM-показників на наступному етапі.

Технічно усі операції інтеграції виконуються мовою R із використанням бібліотеки dplyr. Послідовність викликів inner_join() з відповідними операторами агрегації group_by() та summarise() забезпечує читабельність та повторюваність процесу. Повний код інтеграції даних наведено в Додатку А.

Після завершення всіх join-операцій підсумкова таблиця транзакцій містить 96 469 рядків, що відповідає кількості очищених замовлень, та охоплює 93 349 унікальних клієнтів. Жодних записів не втрачено в результаті інтеграції, що свідчить про коректну реляційну цілісність dataset.

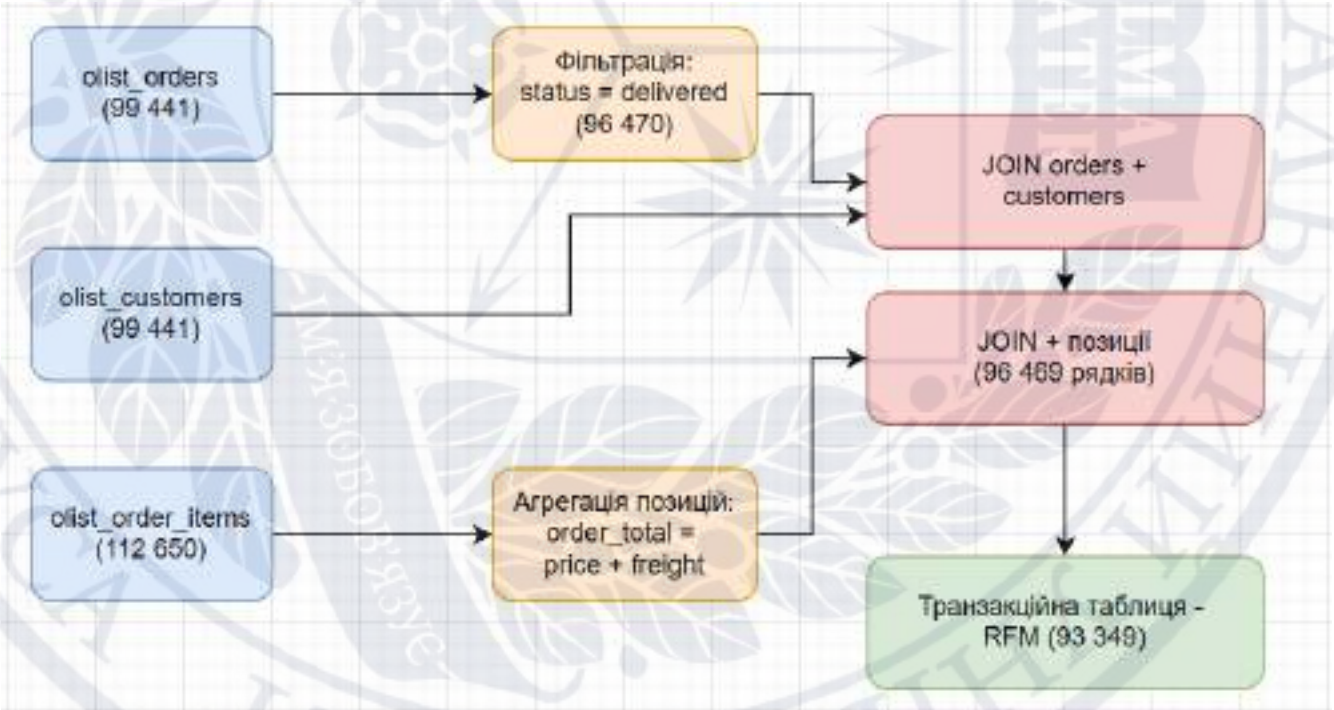


Рисунок 2.1 – Схема інтеграції даних Olist у транзакційний набір

## 2.5. Формування RFM-показників

Формування показників Recency, Frequency, Monetary є важливим етапом дослідження, оскільки саме ці змінні становлять простір ознак для наступної кластеризації. Обчислення показників виконується на основі підсумкової транзакційної таблиці шляхом групування за полем customer_unique_id.

Першим кроком встановлюється референсна дата аналізу. За методичним стандартом RFM-аналізу референсною датою обирається день, наступний після останньої транзакції в dataset. Для досліджуваного набору максимальна дата покупки становить 29 серпня 2018 року, тому референсна дата встановлена як 30 серпня 2018 року. Такий підхід забезпечує мінімальне значення Recency на рівні 1 дня для найактивніших клієнтів.

Показник Recency (R) для кожного клієнта обчислюється як різниця у днях між референсною датою аналізу та датою його останньої транзакції у наборі. Чим менше значення R, тим актуальніше за часом клієнт здійснював купівлю, і тим вищою є його поточна активність відносно референсної дати.

Показник Frequency (F) визначається як кількість унікальних замовлень клієнта за весь період dataset. Для кожного клієнта значення F дорівнює кількості різних ідентифікаторів order_id, асоційованих з його customer_unique_id. У такий спосіб різні товарні позиції в межах одного замовлення не подвоюють значення F – частота вимірюється саме на рівні транзакцій, а не товарних позицій.

Показник Monetary (M) обчислюється як сумарна вартість усіх замовлень клієнта за період аналізу. Для кожного клієнта значення M дорівнює сумі величин order_total по всіх його замовленнях; величина order_total, у свою чергу, формується як сума цін товарних позицій та вартості доставки в межах одного замовлення. Показник Monetary відображає сукупний грошовий внесок клієнта у виручку підприємства за досліджуваний період.

У результаті виконання групування формується RFM-таблиця обсягом 93 349 рядків (за кількістю унікальних клієнтів) та трьома числовими колонками R, F, M. Кожен рядок цієї таблиці відображає поведінковий профіль одного клієнта, який

характеризується трьома узагальненими показниками його транзакційної активності.

Описова статистика отриманих показників наведена у табл. 2.4. Вже на цьому етапі стає очевидною ключова особливість датасету: переважна більшість клієнтів здійснили лише одне замовлення. Медіанне значення Frequency дорівнює 1, а 75-й перцентиль також дорівнює 1; це означає, що принаймні 75 % клієнтів є «одноразовими» (one-time buyers). Це характерна риса маркетплейс-моделі та країн з ринками, що розвиваються, де лояльна повторна купівля поки не сформувалася як домінуюча модель поведінки.

Таблиця 2.4 – Описова статистика RFM-показників (n = 93 349)

| **Показник** | **Мінімум** | **25-й перц.** | **Медіана** | **75-й перц.** | **Середнє** | **Максимум** |
|---|---|---|---|---|---|---|
| Recency, дні | 1,0 | 114,9 | 219,6 | 346,9 | 238,5 | 714,1 |
| Frequency, шт. | 1 | 1 | 1 | 1 | 1,03 | 15 |
| Monetary, R$ | 9,59 | 63,01 | 107,78 | 182,51 | 165,17 | 13 664,08 |

Період Recency варіюється від 1 до 714 днів, що відображає весь часовий діапазон активності клієнтів у датасеті. Показник Monetary охоплює широкий діапазон значень: від мінімальних замовлень близько R$ 10 до максимальних понад R$ 13 600, з вираженою правосторонньою скошеністю розподілу. Сформована RFM-таблиця є вхідним матеріалом для подальшого розвідувального аналізу та побудови кластерної моделі.

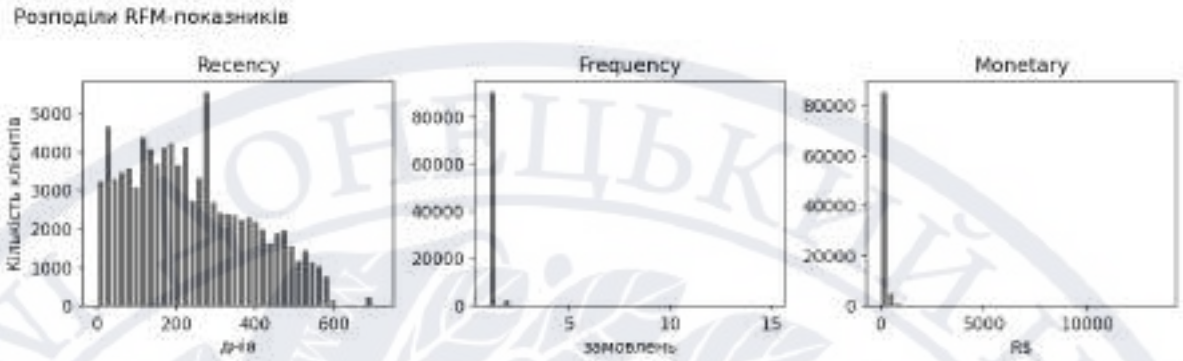


Рисунок 2.2 – Розподіли RFM-показників

### 2.6. Розвідувальний аналіз RFM-показників

Розвідувальний аналіз даних (Exploratory Data Analysis, EDA) проводиться з метою кращого розуміння структури RFM-показників, виявлення можливих аномалій та обґрунтування подальших кроків обробки. Аналіз охоплює вивчення розподілів, кореляційних зв'язків та виявлення викидів.

Аналіз форми розподілів свідчить про істотну асиметрію всіх трьох показників. Коефіцієнти асиметрії, обчислені за стандартною формулою як третій центральний момент, нормований на куб стандартного відхилення, наведено у табл. 2.5.

Таблиця 2.5 – Коефіцієнти асиметрії RFM-показників

| **Показник** | **Асиметрія (raw)** | **Асиметрія (log)** | **Інтерпретація** |
|---|---|---|---|
| Recency | 0,45 | -1,20 | Помірно правостороння - лівостороння після log |
| Frequency | 11,09 | 6,52 | Сильно правостороння; залишається після log |
| Monetary | 9,21 | 0,53 | Сильно правостороння - майже симетрична після log |

Найвищі значення асиметрії демонструють показники Frequency та Monetary, що пояснюється природою даних: невелика кількість клієнтів робить дуже багато покупок або витрачає значно більше за середнього клієнта. Для Monetary логарифмічна трансформація ефективно нормалізує розподіл (асиметрія знижується з 9,21 до 0,53). Для Frequency логарифмічна трансформація лише частково знижує асиметрію (з 11,09 до 6,52), оскільки наявність переважної маси значень F=1 не може бути скоригована трансформацією.

Виявлення викидів проводилося шляхом аналізу 99-го перцентиля показника Monetary. 99-й перцентиль становить R$ 1 097,08, тоді як максимальне значення сягає R$ 13 664,08. Це означає, що приблизно 1 % клієнтів характеризуються винятково високими грошовими витратами, які можуть домінувати в розрахунку центроїдів кластерів. Для зниження впливу цих екстремальних значень на якість кластеризації застосовано техніку winsorization – обмеження значень Monetary на рівні 99-го перцентиля. Це дозволяє зберегти цих клієнтів у вибірці без їхнього виключення, одночасно нівелюючи деструктивний вплив на K-means алгоритм, який чутливий до викидів.

Підсумкові висновки EDA: показники R, F, M мають низьку взаємну кореляцію, що робить їх інформативними ознаками для кластеризації; розподіли значень характеризуються правосторонньою асиметрією, що потребує логарифмічної трансформації перед побудовою моделі; показник Frequency має критично низьку дисперсію через домінування значення F=1, що буде враховано при інтерпретації результатів кластеризації; для зниження впливу викидів Monetary застосовано обмеження на рівні 99-го перцентиля.

### 2.7. Підготовка даних до кластеризації: трансформація та масштабування

Алгоритм K-means базується на евклідовій відстані між точками у багатовимірному просторі ознак. Це утворює дві ключові вимоги до вхідних даних:

ознаки мають бути приведені до симетричного або хоча б приблизно нормального розподілу, та повинні мати порівнянні масштаби. Без виконання цих умов алгоритм демонструє нестабільну поведінку та формує кластери, що відображають властивості шкали, а не реальну структуру даних.

Першим кроком підготовки виконується логарифмічна трансформація показників за формулою log(1 + x). Використання саме цього варіанту (а не звичайного логарифма) дозволяє обробляти нульові значення без появи -∞ та має суттєвий ефект для нормалізації правосторонніх розподілів. Як показано у попередньому підрозділі, після трансформації асиметрія Monetary знижується з 9,21 до 0,53, що робить розподіл придатним для застосування алгоритмів, чутливих до форми розподілу.

Другим кроком виконується стандартизація (z-score нормалізація) – приведення значень до розподілу із середнім 0 та стандартним відхиленням 1 за формулою (2.1):

$$Z = \frac{X - \mu}{\sigma} \tag{2.1}$$

де X – вихідне значення досліджуваної ознаки;

μ – середнє значення ознаки;

σ – стандартне відхилення ознаки.

Стандартизація забезпечує рівноважний вплив кожної з трьох RFM-ознак на розрахунок евклідової відстані. Без неї ознака Recency (діапазон у сотні днів) повністю домінувала б над Frequency (діапазон 1–15), що призвело б до фактичної кластеризації лише за параметром давності.

У результаті виконання двох кроків трансформації формується матриця масштабованих ознак X розмірністю 93 349 × 3, у якій кожна колонка має середнє 0 та стандартне відхилення 1. Саме ця матриця подається на вхід алгоритму K-means для побудови кластерної моделі.

### 2.8. Визначення оптимальної кількості кластерів

Алгоритм K-means потребує попереднього задання кількості кластерів K. Оптимальне значення K не може бути визначене автоматично та потребує застосування спеціальних методів обґрунтування. У цій роботі використано комбінацію двох методів: метод ліктя (Elbow method) та коефіцієнт силуету (silhouette coefficient).

Метод ліктя ґрунтується на побудові залежності внутрішньокластерної дисперсії (WCSS, within-cluster sum of squares) від кількості кластерів. Зі збільшенням K значення WCSS монотонно зменшується, оскільки кластери стають дрібнішими та точки розташовуються ближче до своїх центроїдів. Оптимальне K відповідає точці «перегину», після якої подальше збільшення кількості кластерів не призводить до істотного зниження WCSS. Коефіцієнт силуету, як описано у підрозділі 1.6, вимірює якість кластеризації через порівняння середніх відстаней до власного та найближчого сусіднього кластерів.

Обчислення проводилися для діапазону K = 2..7. Результати наведено у табл. 2.6.

Таблиця 2.6 – Результати визначення оптимальної кількості кластерів

| **K** | **WCSS** | **Зниження WCSS, %** | **Коефіцієнт силуету** |
|---|---|---|---|
| 2 | 214 853,5 | – | 0,711 |
| 3 | 129 013,4 | 39,9 % | 0,385 |
| 4 | 82 050,1 | 36,4 % | 0,397 |
| 5 | 67 671,2 | 17,5 % | 0,357 |
| 6 | 57 840,5 | 14,5 % | 0,355 |
| 7 | 50 543,2 | 12,6 % | 0,365 |

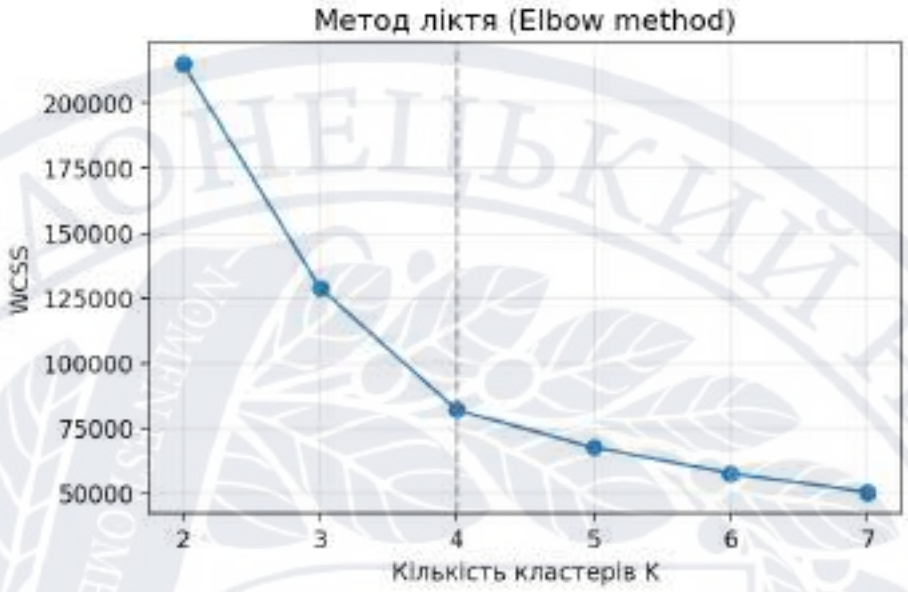


Рисунок 2.3 – Метод ліктя (залежність WCSS від кількості кластерів)

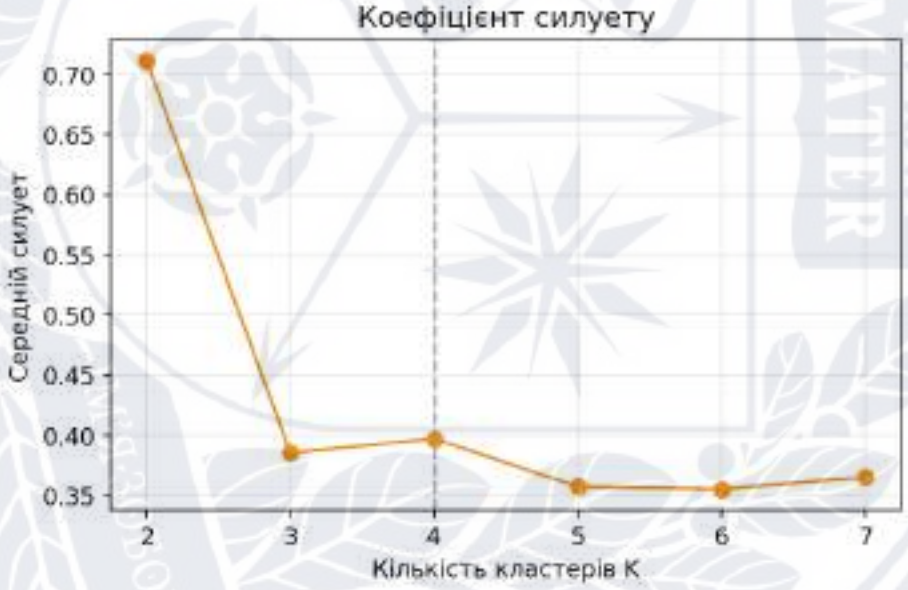


Рисунок 2.4 – Залежність середнього коефіцієнта силуету від кількості кластерів

Аналіз таблиці та рисунків 2.3–2.4 виявляє таку картину. Метод ліктя показує найбільші зниження WCSS при переходах від K=2 до K=3 (-39,9 %) та від K=3 до K=4 (-36,4 %); після K=4 темп зниження істотно скорочується (-17,5 % при переході до K=5 та менше). Це свідчить про «лікоть» приблизно на K=4. Коефіцієнт силуету

має найвище значення при K=2 (0,711), однак двокластерна модель є надто грубою для цілей бізнес-сегментації – вона лише розділяє клієнтів на нещодавніх та давніх, не дозволяючи виділити змістовні поведінкові групи за рівнем витрат та лояльності. Така властивість коефіцієнта силуету – систематичне тяжіння до малих значень K – є відомим обмеженням цієї метрики та потребує комплементарного застосування з методом ліктя. При K=4 силует становить 0,397, що навіть перевищує значення при K=3 (0,385) та є вищим за значення при K=5–6.

З урахуванням компромісу між статистичною якістю кластеризації та змістовністю отриманих сегментів для побудови підсумкової моделі обрано K = 4. Це значення забезпечує одночасно: достатню деталізацію сегментів для практичних маркетингових рішень, високий темп зниження WCSS на попередніх кроках, прийнятний рівень коефіцієнта силуету. Чотирикластерна модель також відповідає поширеній бізнес-практиці виділення сегментів за схемою «нові – лояльні – у зоні ризику – втрачені», що інтуїтивно зрозуміла для прийняття управлінських рішень.

### 2.9. Побудова кластерної моделі методом K-means

Побудова кластерної моделі виконується алгоритмом K-means з кількістю кластерів K = 4. Ініціалізація центроїдів здійснюється через метод k-means++, який забезпечує кращу збіжність порівняно з випадковою ініціалізацією [20]. Для відтворюваності результатів зафіксовано значення випадкового стартового зерна (random seed) на рівні 42. Максимальна кількість ітерацій встановлена на рівні 50, критерій збіжності – зміна координат центроїдів менша за 1e-4 у евклідовій метриці.

Алгоритм досягнув збіжності за обмежену кількість ітерацій. Фінальне значення внутрішньокластерної дисперсії склало WCSS = 82 050,1, а відношення міжкластерної дисперсії до загальної (between_SS / total_SS) становить 0,707, що свідчить про те, що модель пояснює близько 70,7 % загальної варіації даних – це добрий показник для реальних поведінкових даних.

Розміри отриманих кластерів та їхні характеристики на рівні «сирих» (нестандартизованих) RFM-показників подано у табл. 2.7. Кожен рядок таблиці відображає середні та медіанні значення R, F, M по кожному з чотирьох кластерів, абсолютний розмір кластера та частку у вибірці, а також частку у сумарному грошовому обсязі покупок.

Таблиця 2.7 – Профілі отриманих кластерів за RFM-показниками

| Кластер | Клієнтів | Частка, % | R, дні | F | M, R$ | Частка доходу, % |
|---|---|---|---|---|---|---|
| 1 | 16 203 | 17,36 | 42,9 | 1,00 | 135,05 | 14,19 |
| 2 | 2 801 | 3,00 | 220,8 | 2,11 | 308,53 | 5,60 |
| 3 | 41 812 | 44,79 | 288,3 | 1,00 | 67,66 | 18,35 |
| 4 | 32 533 | 34,85 | 273,3 | 1,00 | 293,15 | 61,85 |
| Всього | 93 349 | 100,00 | 238,5 | 1,03 | 165,17 | 100,00 |

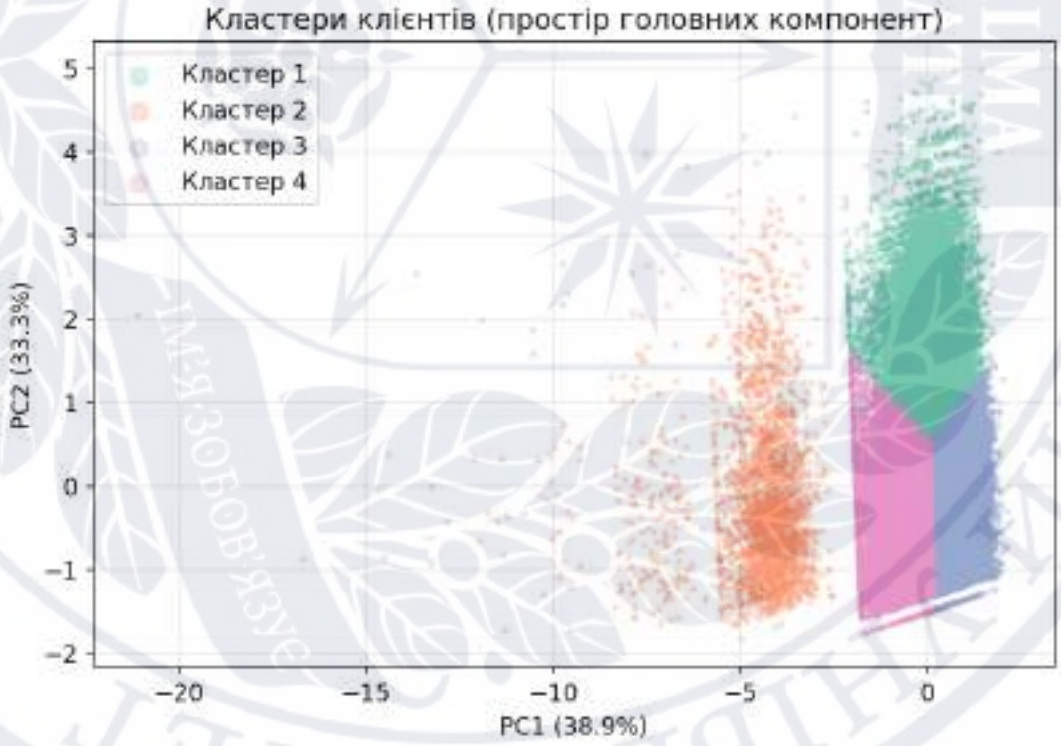


Рисунок 2.5 – Візуалізація кластерів у просторі перших двох головних компонент

Розподіл клієнтів демонструє виражену асиметрію – найбільший кластер 3 охоплює 44,79 % клієнтської бази, найменший кластер 2 – лише 3,00 %. Аналогічна асиметрія спостерігається у структурі сумарного доходу: кластер 4 з 34,85 % клієнтів генерує 61,85 % сумарного грошового обсягу покупок (закон Парето у дії), тоді як кластер 3 з 44,79 % клієнтів генерує лише 18,35 % доходу. Це підкреслює важливість якісної сегментації – однакові маркетингові зусилля для всіх клієнтів економічно недоцільні.

Візуалізація кластерів у просторі перших двох головних компонент (рис. 2.5), на які припадає 72,2 % сумарної дисперсії (38,9 % та 33,3 % відповідно), наочно демонструє відокремленість кластера 2 (повторні клієнти) від решти груп уздовж першої компоненти, а також поділ масиву одноразових клієнтів на три підгрупи за рівнем витрат та давністю.

Технічна реалізація алгоритму виконана мовою R із використанням стандартної функції kmeans() пакета stats та допоміжного пакета factoextra для візуалізації. Час обчислення на комп'ютері персонального використання середньої цінової категорії не перевищує 30 секунд, що підтверджує практичну застосовність моделі.

### 2.10. Оцінювання якості кластеризації та інтерпретація сегментів

Оцінювання якості отриманої кластерної моделі здійснюється за двома вимірами: статистичної якості (через метрики кластеризації) та змістовної інтерпретованості (через бізнес-значущість виявлених сегментів).

З погляду статистичної якості модель характеризується значенням WCSS = 82 050,1, коефіцієнтом силуету 0,397 та відношенням between_SS / total_SS = 0,707. Значення силуету в межах 0,3–0,5 у літературі прийнято інтерпретувати як «помірно якісну» структуру кластеризації – кластери розрізняються між собою, але мають деяке перекриття на межах [23]. Для реальних e-commerce даних із

вираженою асиметрією розподілів такий рівень якості є очікуваним та практично прийнятним.

Змістовна інтерпретація сегментів базується на порівнянні їхніх центроїдів та маркетинговій типології, описаній у підрозділі 1.7. На основі профілів з табл. 2.8 кожному кластеру присвоєно бізнес-назву та сформульовано характеристики цільового сегмента. Узагальнена інтерпретація сегментів подана у табл. 2.9.

Таблиця 2.8 – Бізнес-інтерпретація отриманих сегментів

| **Кластер** | **Назва сегмента** | **Профіль** | **Маркетингова стратегія** |
|---|---|---|---|
| 1 | Recent / New Customers | Нещодавня купівля (R=43), один чек R$ 135 | Конверсія в лояльних, welcome-кампанії, повторний контакт |
| 2 | Loyal / Champions | Повторні клієнти (F=2,11), середній чек R$ 309, проміжна давність | Утримання, програми лояльності, ексклюзивні пропозиції |
| 3 | Lost / Low-Value | Давня купівля (R=288), низький чек R$ 68 | Мінімальні маркетингові інвестиції, можливе виключення |
| 4 | Hibernating High-Value | Давня купівля (R=273), високий чек R$ 293, генерує 62 % доходу | Реактивація, персоналізовані пропозиції, win-back кампанії |

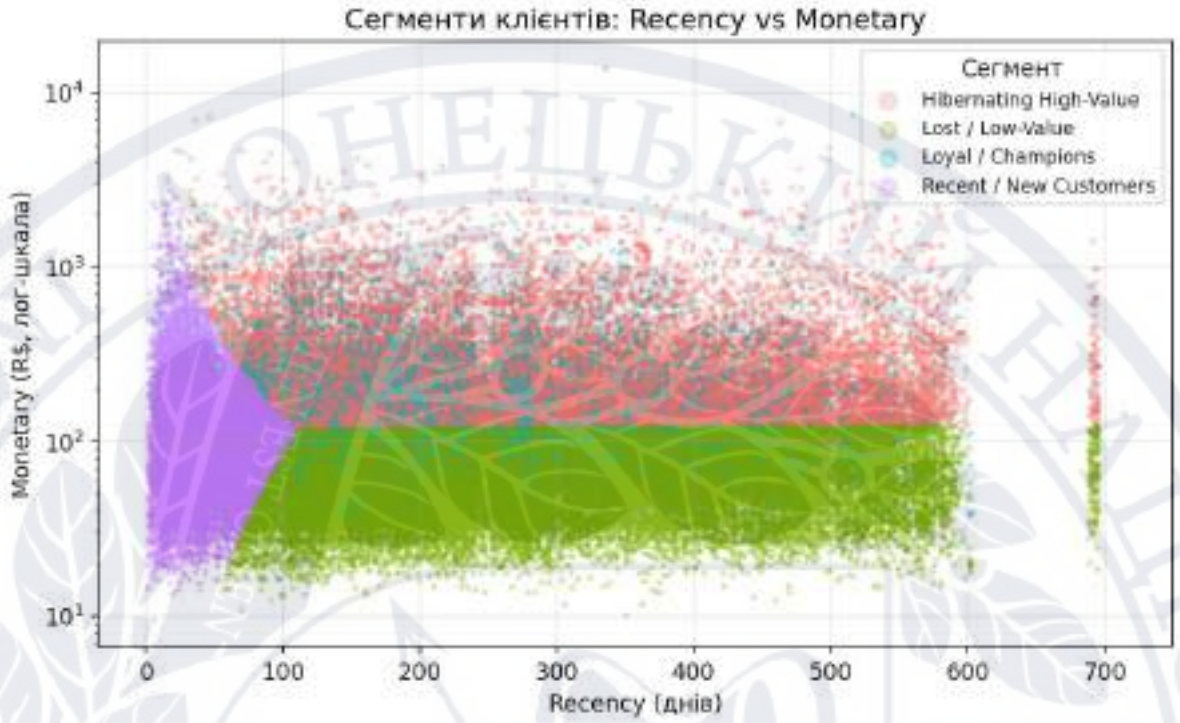


Рисунок 2.6 – Розподіл клієнтів за сегментами у координатах Recency–Monetary

Найвагоміший практичний інсайт стосується кластера 4 (Hibernating High-Value). Цей сегмент включає 32 533 клієнти (34,85 % бази), які витрачали в середньому R$ 293 за замовлення та сукупно сформували 61,85 % сумарного доходу. Однак ці клієнти не здійснювали покупок понад дев'ять місяців (середнє Recency 273 дні), що означає високий ризик їхньої повної втрати. Спрямування маркетингових ресурсів на реактивацію саме цього сегмента має максимальний потенційний поворот інвестицій, оскільки історичні витрати свідчать про їхню готовність робити покупки на значні суми.

Кластер 1 (Recent / New Customers, 17,36 % бази) є стратегічно важливим з іншої причини – це найсвіжіший потік клієнтів, які купили один раз. Конверсія їх у повторних покупців може якісно змінити структуру клієнтської бази та зменшити частку «одноразових» транзакцій. Маркетингова стратегія тут має фокусуватися на формуванні повторного контакту, рекомендаціях супутніх товарів та програмах лояльності.

Кластер 2 (Loyal / Champions, 3,00 %) – найменший за розміром, однак найбільш стабільний з погляду регулярності покупок (єдиний сегмент із середнім Frequency понад 2). Це ядро лояльної клієнтської бази, для якого виправдано впроваджувати VIP-програми, персоналізовані рекомендації та ексклюзивні умови. Кластер 3 (Lost / Low-Value, 44,79 %) – це однократні низькочекові покупців, інвестиції в реактивацію яких економічно недоцільні; такі клієнти можуть бути виключені з активних маркетингових кампаній з метою оптимізації витрат.

Аналізуючи отримані результати у ширшому контексті, важливо зазначити, що структура клієнтської бази Olist відображає типову модель ринку, що розвивається: висока частка одноразових покупців, низький рівень повторних транзакцій, концентрація доходу у вузькому сегменті. Це є нормальним для маркетплейсу країни Латинської Америки і не свідчить про методологічні проблеми кластеризації. Скоріше, виявлена структура є цінним усвідомленням, що чітко вказує на стратегічні пріоритети – побудова лояльності та реактивація неактивних клієнтів є центральними завданнями маркетингу для платформ такого типу.

Сформована кластерна модель є основою для майбутньої практичної реалізації аналітичної системи, що детально описано у Розділі 3. Зокрема, отриманий набір сегментів буде використано для побудови інтерактивного дашборду в Power BI, який дозволить маркетинговому персоналу візуально досліджувати кожну з виявлених груп клієнтів та формувати адаптовані стратегії взаємодії.

**Висновки до розділу 2**

У другому розділі розроблено модель сегментації клієнтів на основі реального набру даних Brazilian E-Commerce Public Dataset by Olist. Проведено порівняльний аналіз кандидатних датасетів, обґрунтовано вибір Olist на основі поєднання критеріїв реальності даних, керованого обсягу та сучасності збору.

Реалізовано повний цикл попередньої обробки даних, що охопив фільтрацію замовлень за статусом доставки, обробку пропущених значень у часових мітках, перетворення типів та перевірку цілісності зв'язків між таблицями. Виконано інтеграцію даних через послідовність join-операцій між таблицями замовлень, клієнтів та позицій, що сформувала єдину транзакційну таблицю придатну для подальшого аналізу.

На основі очищеного транзакційного набору сформовано показники Recency, Frequency, Monetary для кожного клієнта. Виконано розвідувальний аналіз отриманих показників, що виявив виражену асиметрію розподілів Monetary та Frequency.

Здійснено підготовку даних до кластеризації шляхом логарифмічної трансформації та z-score стандартизації, що дозволило нормалізувати розподіли та забезпечити рівноважний вплив усіх ознак на евклідову відстань. Визначення оптимальної кількості кластерів виконано комбінацією методу ліктя та коефіцієнта силуету, на основі чого обрано K = 4 як значення, що забезпечує компроміс між статистичною якістю та змістовністю сегментів.

Проведено бізнес-інтерпретацію отриманих сегментів та сформульовано адаптовані маркетингові стратегії для кожного з них. Сформована кластерна модель становить основу для розробки аналітичної системи у Розділі 3.

# РОЗДІЛ 3. РОЗРОБКА ТА РЕАЛІЗАЦІЯ АНАЛІТИЧНОЇ СИСТЕМИ СЕГМЕНТАЦІЇ КЛІЄНТІВ

## 3.1. Архітектура аналітичної системи сегментації

Розроблена аналітична система сегментації клієнтів є інтегрованим рішенням, що поєднує статистичну обробку даних мовою R з інтерактивною візуалізацією засобами Microsoft Power BI. Архітектура системи побудована за принципом розподілу відповідальності між компонентами: обчислювально складні етапи статистичного навчання реалізовано у спеціалізованому середовищі R, а представлення результатів кінцевому користувачеві – у бізнес-аналітичній платформі Power BI. Такий підхід дозволяє поєднати потужність статистичних методів із доступністю інтерактивної аналітики для нетехнічних користувачів.

Загальна архітектура системи складається з шести логічних компонентів. Першим є джерело даних – реальний транзакційний набір Brazilian E-Commerce Public Dataset by Olist у форматі дев'яти CSV-таблиць. Другим компонентом є шар обробки, реалізований мовою R у середовищі RStudio, що виконує завантаження, очищення, інтеграцію даних, формування RFM-показників та кластеризацію методом K-means. Третім компонентом є проміжний файл обміну – згенерований R-скриптом CSV-файл із результатами сегментації. Четвертим є BI-шар (Power BI), що імпортує дані, формує семантичну модель та обчислювані метрики. П'ятим компонентом є аналітичний дашборд, що візуалізує результати. Шостим – кінцевий користувач (клієнт, маркетолог, менеджер), який взаємодіє з дашбордом для прийняття рішень.

Архітектуру системи у графічному вигляді подано на рис. 3.1.

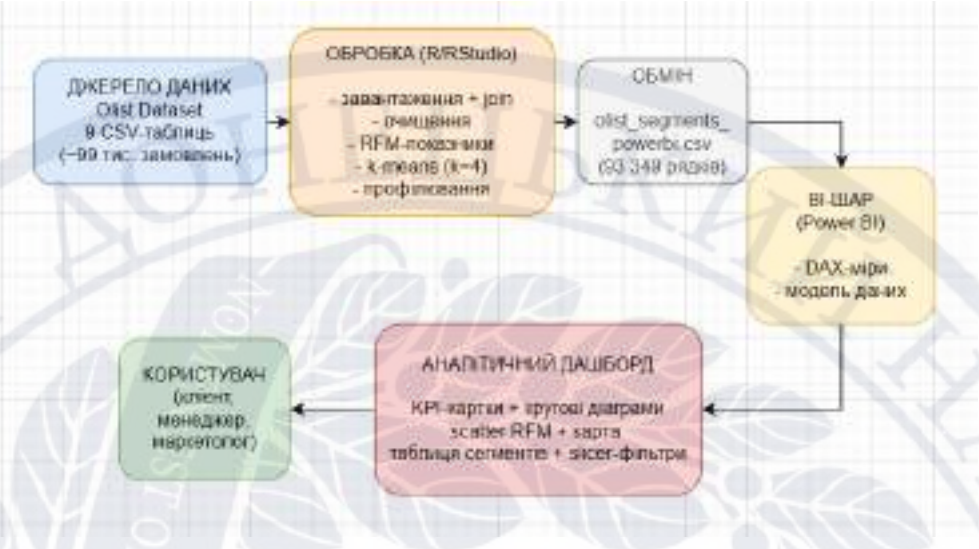


Рисунок 3.1 – Архітектура аналітичної системи сегментації клієнтів

Ключовою перевагою такої архітектури є слабка зв'язаність компонентів (loose coupling). Шар обробки даних та шар візуалізації обмінюються інформацією через стандартний CSV-формат, що робить кожен компонент незалежним та замінним. За потреби R-скрипт може бути замінений на Python-реалізацію, а Power BI – на іншу BI-платформу без зміни загальної логіки системи. Окрім цього, така архітектура забезпечує відтворюваність: повторний запуск R-скрипту на оновлених даних автоматично оновлює вхідний файл для Power BI, що дозволяє регулярно актуалізувати сегментацію.

### 3.2. Обґрунтування вибору мови R та середовища розробки

Для реалізації статистичної частини системи обрано мову програмування R [37]– спеціалізоване середовище для статистичних обчислень та аналізу даних. Вибір R обґрунтований кількома чинниками. По-перше, R має вбудовану підтримку статистичних методів, зокрема функцію kmeans() у базовому пакеті stats, що не потребує додаткових залежностей для кластеризації. По-друге, екосистема R містить розвинені пакети для роботи з даними (tidyverse) та кластерного аналізу (cluster, factoextra, NbClust), які покривають усі етапи дослідження. По-третє, R є відкритим та безкоштовним програмним забезпеченням, що відповідає орієнтації

системи на сегмент малого та середнього бізнесу. По-четверте, R забезпечує повну відтворюваність обчислень за рахунок фіксації випадкового зерна (set.seed), що критично важливо для академічного дослідження.

Як середовище розробки використано RStudio Desktop – інтегроване середовище розробки (IDE), що надає зручний інтерфейс для написання та виконання R-коду, перегляду даних, побудови графіків та управління проєктом. RStudio підтримує покрокове виконання скриптів, інтерактивний перегляд змінних та вбудовану панель візуалізації, що пришвидшує процес розробки та налагодження аналітичних скриптів.

Структурно R-проєкт організовано як єдиний скрипт, поділений на логічні секції відповідно до етапів обробки: підключення бібліотек, завантаження даних, очищення, інтеграція, формування RFM, кластеризація, профілювання та експорт. Такий підхід забезпечує читабельність коду та відповідає принципу послідовного виконання аналітичного конвеєра. Повний код реалізації наведено у Додатку А.

### 3.3. Опис використаних бібліотек R

Реалізація системи спирається на набір спеціалізованих R-пакетів, кожен з яких виконує визначену роль у конвеєрі обробки. Перелік використаних бібліотек та їхнє призначення подано у табл. 3.1.

Таблиця 3.1 – Використані бібліотеки R та їхнє призначення

| Бібліотека | Призначення в системі |
|---|---|
| tidyverse | Метапакет, що включає dplyr (маніпуляції даними), readr (читання CSV), tidyr (трансформація), ggplot2 (візуалізація) Детальний опис екосистеми tidyverse наведено у [38] |
| dplyr | Операції фільтрації, групування (group_by), агрегації (summarise) та об'єднання таблиць (join) |
| lubridate | Парсинг та обробка часових міток (ymd_hms), обчислення різниці дат (difftime) для показника Recency |

| | |
|---|---|
| readr | Швидке читання CSV-файлів (read_csv) та запис результатів (write_csv) |
| cluster | Обчислення коефіцієнта силуету (функція silhouette) для оцінювання якості кластеризації |
| factoextra | Візуалізація результатів кластеризації (fviz_cluster) у просторі головних компонент. Методологічну основу застосування цих пакетів для кластерного аналізу подано в [39] |
| ggplot2 | Побудова графіків розподілів, методу ліктя, коефіцієнта силуету та діаграм розсіювання. Детальний опис граматики графіків ggplot2 представлено в [40]. |
| stats (базовий) | Реалізація алгоритму K-means (функція kmeans) та стандартизації даних (scale) |

Окремо слід відзначити, що алгоритм K-means реалізовано через стандартну функцію kmeans() базового пакета stats, що не потребує встановлення додаткових залежностей. Це підвищує переносимість коду та спрощує його запуск на будь-якому комп'ютері з установленим R. Логарифмічна трансформація та z-score стандартизація реалізовані вбудованими функціями log1p() та scale() відповідно.

### 3.4. Реалізація модуля завантаження та інтеграції даних

Модуль завантаження та інтеграції даних відповідає за читання вихідних CSV-таблиць, їхнє очищення та об'єднання у єдиний транзакційний набір. Реалізація починається з підключення бібліотек, налаштування робочої директорії та фіксації випадкового зерна для відтворюваності. Завантаження таблиць виконується функцією read_csv() з параметром show_col_types = FALSE для пригнічення службових повідомлень.

Очищення даних реалізовано через ланцюг операцій dplyr. Фільтрація доставлених замовлень та видалення записів із пропущеними датами доставки виконуються послідовними викликами filter(), як показано у фрагменті коду нижче:

```
orders_clean <- orders %>%
  filter(order_status == "delivered") %>%
  filter(!is.na(order_delivered_customer_date))

orders_clean <- orders_clean %>%
  mutate(order_purchase_timestamp =
           ymd_hms(order_purchase_timestamp, quiet = TRUE))
```

Інтеграція даних реалізована послідовністю join-операцій. Спочатку виконується агрегація позицій замовлень до рівня окремого замовлення з обчисленням сумарної вартості, після чого результат об'єднується з таблицями замовлень та клієнтів через спільні ключі order_id та customer_id. Реалізацію агрегації позицій наведено нижче:

```
order_totals <- items %>%
  group_by(order_id) %>%
  summarise(
    order_total_price   = sum(price),
    order_total_freight = sum(freight_value),
    n_items             = n(),
    .groups = "drop"
  ) %>%
  mutate(order_total = order_total_price + order_total_freight)

df <- orders_clean %>%
  inner_join(customers, by = "customer_id") %>%
  inner_join(order_totals, by = "order_id")
```

У результаті виконання модуля формується єдина транзакційна таблиця df обсягом 96 469 рядків, що охоплює 93 349 унікальних клієнтів. Повний код модуля наведено у Додатку А.

### 3.5. Реалізація модуля формування RFM-показників

Модуль формування RFM-показників реалізує обчислення трьох поведінкових метрик для кожного клієнта на основі інтегрованої транзакційної таблиці. Першим кроком встановлюється референсна дата аналізу як день,

наступний після останньої транзакції в наборі. Обчислення показників виконується через групування за ідентифікатором клієнта customer_unique_id з агрегацією відповідних значень:

```
ref_date <- max(df$order_purchase_timestamp, na.rm = TRUE) + days(1)

rfm <- df %>%
  group_by(customer_unique_id) %>%
  summarise(
    Recency   = as.numeric(difftime(ref_date,
                  max(order_purchase_timestamp), units = "days")),
    Frequency = n_distinct(order_id),
    Monetary  = sum(order_total),
    .groups   = "drop"
  )
```

Показник Recency обчислюється як різниця в днях між референсною датою та датою останньої транзакції клієнта за допомогою функції difftime(). Показник Frequency визначається через n_distinct() – кількість унікальних замовлень. Показник Monetary обчислюється як сума вартостей усіх замовлень клієнта. Розвідувальний аналіз отриманих показників реалізовано через обчислення коефіцієнтів асиметрії та кореляційної матриці, а також побудову графіків розподілів засобами ggplot2. Повний код модуля наведено у Додатку А.

### 3.6. Реалізація модуля кластеризації та експорту результатів

Модуль кластеризації реалізує підготовку даних, побудову K-means моделі та експорт результатів для Power BI. Підготовка даних включає обмеження викидів показника Monetary на рівні 99-го перцентиля, логарифмічну трансформацію та z-score стандартизацію. Важливим елементом захисного програмування є очищення матриці ознак від можливих пропущених значень перед запуском алгоритму:

```
m_cap <- quantile(rfm$Monetary, 0.99)
rfm <- rfm %>% mutate(Monetary_capped = pmin(Monetary, m_cap))

rfm <- rfm %>% mutate(
  log_R = log1p(Recency),
  log_F = log1p(Frequency),
  log_M = log1p(Monetary_capped)
)

X <- scale(rfm[, c("log_R", "log_F", "log_M")])
X <- X[complete.cases(X), ]
```

Побудова кластерної моделі виконується функцією kmeans() з параметрами centers = 4 (кількість кластерів, обґрунтована у підрозділі 2.8), nstart = 25 (кількість випадкових ініціалізацій для пошуку кращого рішення) та фіксованим зерном set.seed(42). Експорт результатів формує підсумкову таблицю, що об'єднує RFM-показники, мітки кластерів, бізнес-назви сегментів та географічні атрибути:

```
set.seed(42)
km_final <- kmeans(X, centers = 4, nstart = 25, iter.max = 50)
rfm$cluster <- km_final$cluster

powerbi_export <- rfm %>%
  left_join(geo, by = "customer_unique_id") %>%
  select(customer_unique_id, Recency, Frequency, Monetary,
         cluster, segment, customer_state, customer_city)

write_csv(powerbi_export, "output/olist_segments_powerbi.csv")
```

Результатом виконання модуля є файл olist_segments_powerbi.csv обсягом 93 349 рядків, що містить повну сегментацію клієнтської бази та слугує вхідним джерелом для Power BI. Повний код модуля кластеризації наведено у Додатку А.

### 3.7. Проєктування аналітичного дашборду в Power BI

Проєктування дашборду здійснювалось із орієнтацією на потреби кінцевого користувача та відповідно до офіційних рекомендацій Microsoft [41]– менеджера/маркетолога, який потребує швидкого огляду стану клієнтської бази та

можливості деталізованого аналізу окремих сегментів. Концепція дашборду базується на принципах ефективної візуалізації даних: ключові показники розміщуються у верхній частині для миттєвого сприйняття, а деталізовані візуалізації – нижче, з можливістю інтерактивної фільтрації.

Дашборд організовано у вигляді двох взаємопов'язаних сторінок, що відповідають різним рівням аналізу. Перша сторінка – «Огляд» – призначена для керівного складу та надає узагальнену картину стану клієнтської бази. Друга сторінка – «Аналітика RFM» – орієнтована на поглиблений аналіз сегментів та демонструє деталізовані поведінкові показники. Такий поділ відповідає принципу прогресивного розкриття інформації (progressive disclosure): користувач спершу отримує загальну картину, а за потреби переходить до деталізації.

Сторінку «Огляд» побудовано за принципом «згори вниз, від загального до часткового». Верхній рівень формують KPI-картки із чотирма ключовими метриками: загальний дохід, кількість клієнтів, кількість замовлень та середній чек. Середній рівень містить кругову діаграму розподілу доходу за сегментами та порівняльну стовпчасту діаграму частки клієнтів і частки доходу. Нижній рівень формує географічна карта розподілу доходу за штатами Бразилії. Сторінку «Аналітика RFM» формують матриця-профіль сегментів з умовним форматуванням, діаграма розсіювання у просторі Recency–Monetary та стовпчаста діаграма середнього чека за сегментами. Кожну сторінку доповнено зрізом (slicer) за сегментами, що забезпечує інтерактивну фільтрацію усіх візуалізацій.

Концептуальну структуру дашборду, узгоджену з підходом, апробованим у науковій публікації автора [28], розширено для відображення результатів кластерної сегментації. Якщо у початковій публікації використовувалася спрощена схема поділу на сегменти High/Medium/Low, то в розробленій системі дашборд відображає повноцінні RFM-кластери, отримані методом K-means. Для забезпечення професійного та цілісного візуального стилю до дашборду застосовано єдину тему оформлення з узгодженою палітрою, у якій кольори

сегментів несуть семантичне навантаження: зелений – лояльні клієнти, синій – нові, помаранчевий – цінні клієнти у зоні ризику, сірий – втрачені.

### 3.8. Реалізація дашборду: модель даних та DAX-метрики

Реалізація дашборду розпочинається з імпорту файлу olist_segments_powerbi.csv у Power BI через інструмент Power Query (Get Data → Text/CSV) [42]. На етапі імпорту виконується перевірка та коригування типів даних: оскільки регіональні налаштування системи використовують кому як десятковий роздільник, числові показники Recency та Monetary потребували приведення до числового типу. Для коректного перетворення створено обчислювані стовпці засобами DAX-функції VALUE(), що інтерпретує текстові значення з урахуванням культури моделі (en-US):

```
Recency_num  = VALUE('olist_segments_powerbi'[Recency])
Monetary_num = VALUE('olist_segments_powerbi'[Monetary])
```

Згідно з офіційною документацією [43], на основі підготовленої моделі даних створено набір DAX-метрик (measures), що забезпечують обчислення ключових аналітичних показників. Перелік розроблених метрик та їхні формули подано у табл. 3.2.

Таблиця 3.2 – Розроблені DAX-метрики аналітичної моделі

| Метрика | DAX-формула |
|---|---|
| Total Customers | `DISTINCTCOUNT([customer_unique_id])` |
| Total Revenue | `SUM([Monetary_num])` |
| Total Orders | `SUM([Frequency])` |
| Avg Order Value | `DIVIDE([Total Revenue], [Total Orders])` |
| Avg Revenue per Customer | `DIVIDE([Total Revenue], [Total Customers])` |
| Avg Recency (days) | `AVERAGE([Recency_num])` |
| Avg Frequency | `AVERAGE([Frequency])` |

| | |
|---|---|
| Revenue Share % | `DIVIDE([Total Revenue], CALCULATE([Total Revenue], ALL(...)))` |
| Customer Share % | `DIVIDE([Total Customers], CALCULATE([Total Customers], ALL(...)))` |
| Repeat Customers | `CALCULATE([Total Customers], [Frequency] > 1)` |
| Repeat Rate % | `DIVIDE([Repeat Customers], [Total Customers])` |

Метрики Revenue Share % та Customer Share % використовують функцію CALCULATE з модифікатором ALL для обчислення частки сегмента у загальному обсязі незалежно від застосованих фільтрів, що дозволяє коректно відображати внесок кожного сегмента. Метрики Repeat Customers та Repeat Rate % обчислюють кількість та частку клієнтів із повторними покупками. Формати метрик налаштовано відповідно до їхнього змісту: грошові показники відображаються з префіксом «R$», частки – у відсотковому форматі.

### 3.9. Опис візуалізацій дашборду

Аналітичний дашборд містить набір взаємопов'язаних візуалізацій, розподілених на двох сторінках. Загальний вигляд сторінки «Огляд» подано на рис. 3.2.

KPI-картки у верхній частині сторінки відображають чотири ключові показники діяльності: загальний дохід (R$ 15 млн), кількість клієнтів (93 349), кількість замовлень (96 469) та середній чек (R$ 159,83). Ці метрики забезпечують миттєвий огляд масштабу клієнтської бази та обсягів продажів.

Кругова діаграма розподілу доходу за сегментами наочно демонструє домінування сегмента Hibernating High-Value, який генерує 61,85 % сумарного доходу. Така візуалізація відразу привертає увагу до ключового висновку дослідження – концентрації доходу у вузькому сегменті клієнтів, що давно не здійснювали покупок.

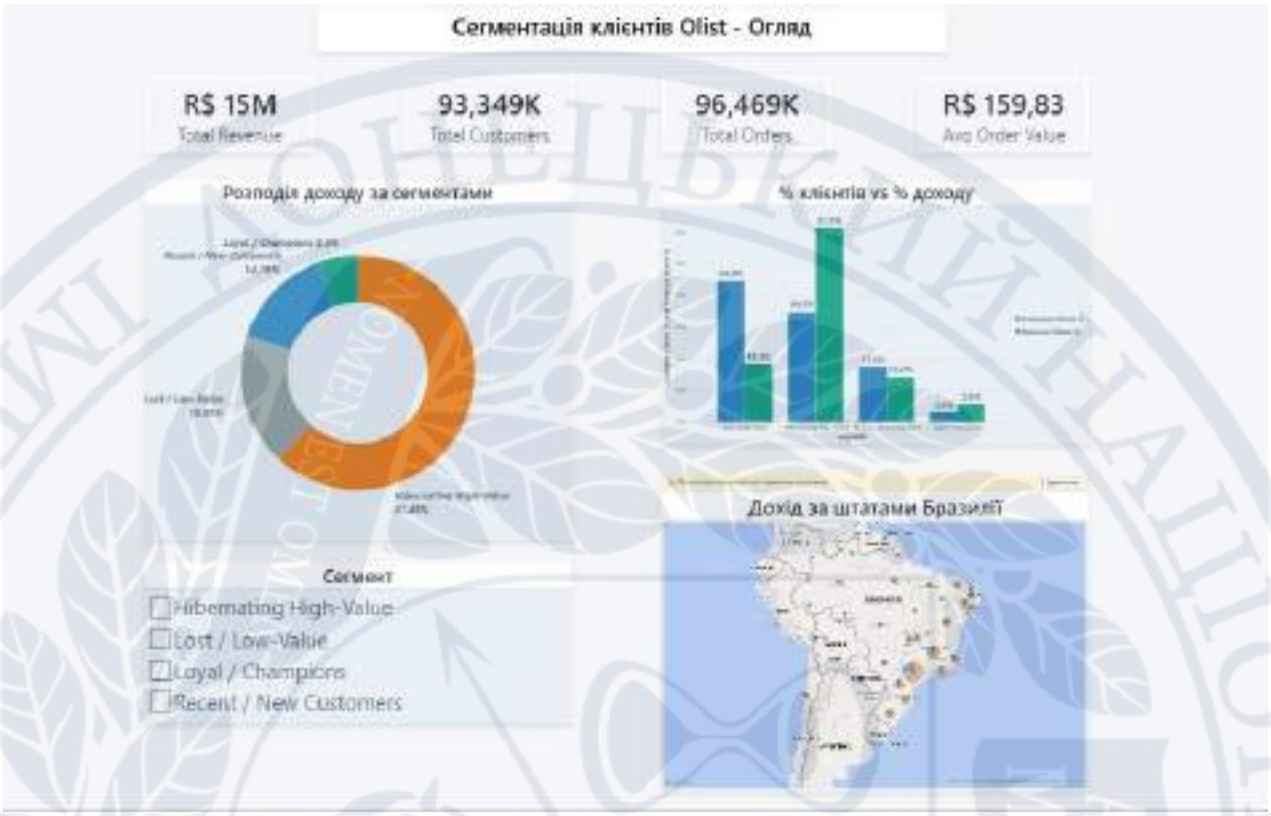


Рисунок 3.2 – Сторінка «Огляд» аналітичного дашборду сегментації клієнтів

Особливе аналітичне значення має порівняльна стовпчаста діаграма частки клієнтів та частки доходу за сегментами (рис. 3.3). Вона візуалізує диспропорцію між кількісним та вартісним внеском кожного сегмента: сегмент Hibernating High-Value охоплює 34,9 % клієнтів, але генерує 61,9 % доходу, тоді як сегмент Lost / Low-Value, навпаки, становить 44,8 % клієнтів, але дає лише 18,3 % доходу. Ця діаграма є графічним відображенням принципу Парето та найпереконливіше обґрунтовує необхідність диференційованого підходу до сегментів.

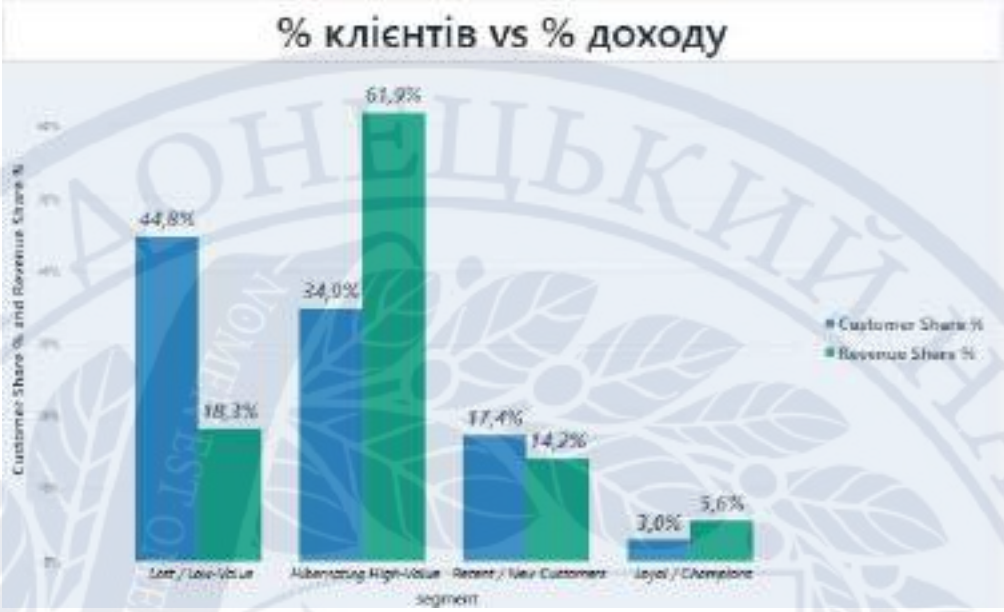


Рисунок 3.3 – Порівняння частки клієнтів та частки доходу за сегментами


Рисунок 3.4 – Розподіл доходу за штатами Бразилії

Географічна карта розподілу доходу за штатами Бразилії (рис. 3.4) відображає просторову структуру клієнтської бази з використанням реальних географічних координат штатів. Аналіз карти підтверджує домінування південно-східного регіону: штат São Paulo генерує найбільший дохід, за ним слідують Rio de Janeiro

та Minas Gerais. Така візуалізація корисна для планування регіональних маркетингових кампаній та логістики.

Друга сторінка дашборду – «Аналітика RFM» – призначена для поглибленого аналізу поведінкових характеристик сегментів. Її загальний вигляд подано на рис. 3.5.

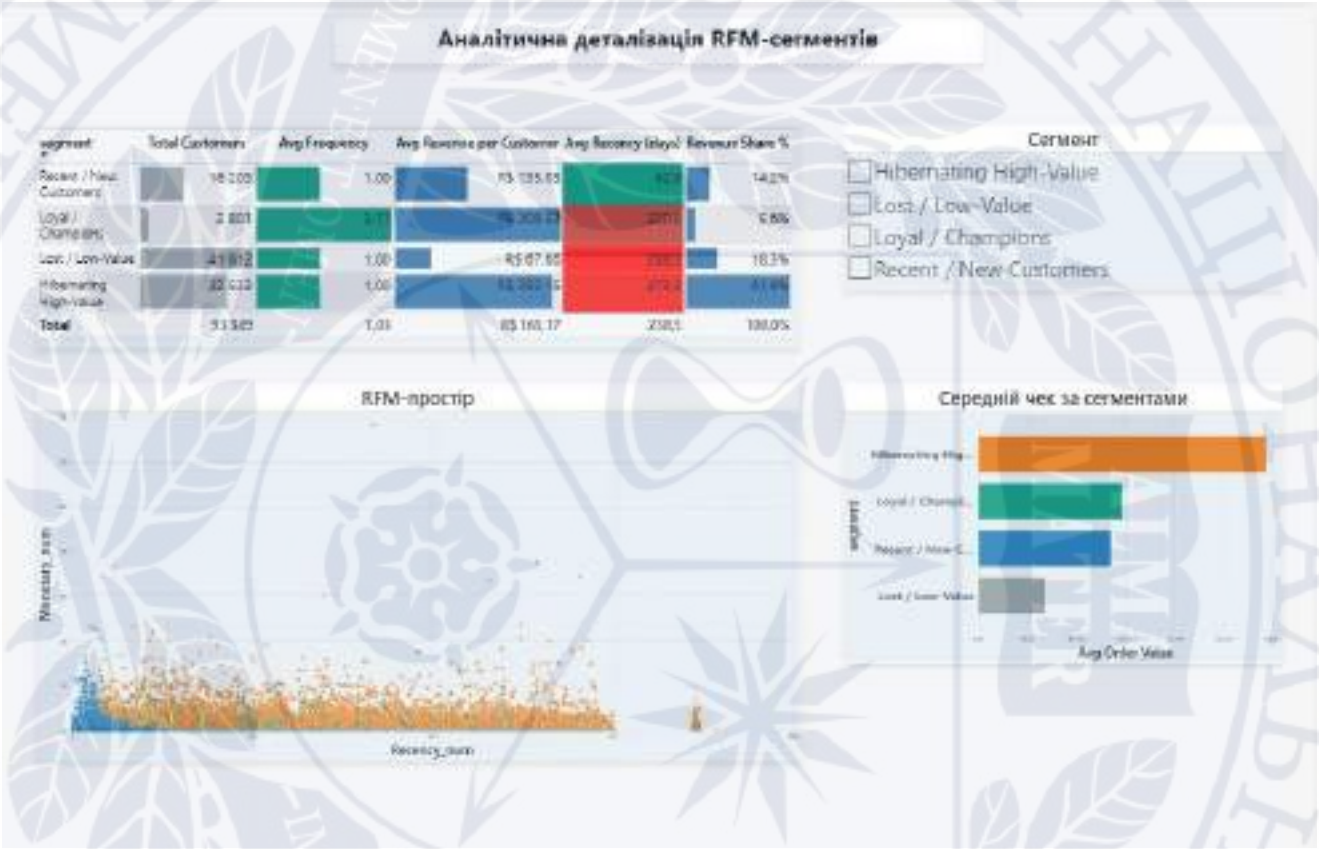


Рисунок 3.5 – Сторінки «Аналітика RFM» аналітичного дашборду

Центральним елементом цієї сторінки є матриця-профіль сегментів з умовним форматуванням (рис. 3.6). Матриця подає для кожного сегмента п'ять показників: кількість клієнтів, середню частоту, середній дохід на клієнта, середню давність та частку доходу. Застосування умовного форматування перетворює таблицю на «теплову карту»: показник давності (Recency) відображається колірною шкалою від зеленого (свіжі клієнти) до червоного (сплячі), а решта показників – горизонтальними смужками (data bars), довжина яких пропорційна значенню.

Такий підхід дозволяє одним поглядом оцінити стан кожного сегмента: наприклад, сегмент Recent / New Customers має зелену клітинку давності (42,9 дня), тоді як Hibernating High-Value та Lost / Low-Value – червоні (273 та 288 днів відповідно), при цьому Hibernating демонструє найдовшу смужку частки доходу (61,9 %).

| segment | Total Customers | Avg Frequency | Avg Revenue per Customer | Avg Recency (days) | Revenue Share % |
|---|---|---|---|---|---|
| Recent / New Customers | 16 203 | 1,00 | R$ 135,05 | 42,9 | 14,2% |
| Loyal / Champions | 2 801 | 2,11 | R$ 308,53 | 220,8 | 5,6% |
| Lost / Low-Value | 41 812 | 1,00 | R$ 67,66 | 288,3 | 18,3% |
| Hibernating High-Value | 32 533 | 1,00 | R$ 293,15 | 273,3 | 61,9% |
| **Total** | 93 349 | 1,03 | R$ 165,17 | 238,5 | 100,0% |

Рисунок 3.6 – Матриця-профіль сегментів з умовним форматуванням

Діаграма розсіювання у просторі Recency–Monetary візуалізує позиціонування кожного клієнта у двовимірному просторі поведінкових показників із кольоровим кодуванням за сегментами. Вона дозволяє візуально оцінити структуру сегментів: видно щільне скупчення нових клієнтів (синій колір) у зоні низької давності та розсіяний масив сплячих клієнтів (помаранчевий) уздовж осі давності. Стовпчаста діаграма середнього чека за сегментами доповнює аналіз, демонструючи, що найвищий середній чек мають сегменти Hibernating High-Value та Loyal / Champions, а найнижчий – Lost / Low-Value.

### 3.10. Інтерактивність та сценарії використання дашборду

Ключовою перевагою реалізованого дашборду є інтерактивність, що забезпечується механізмом перехресної фільтрації (cross-filtering) Power BI. Усі візуалізації дашборду пов'язані між собою: вибір значення на одній візуалізації автоматично фільтрує дані на всіх інших. Окрім цього, на дашборді розміщено зріз (slicer) за сегментами, що дозволяє користувачеві сфокусувати аналіз на конкретній групі клієнтів. Зрізи на обох сторінках синхронізовано за допомогою функції синхронізації зрізів (synced slicers): обраний користувачем сегмент застосовується

одночасно і до сторінки «Огляд», і до сторінки «Аналітика RFM», що забезпечує цілісність аналізу під час переходу між сторінками.

Розглянемо типові сценарії використання дашборду. Перший сценарій – аналіз окремого сегмента. Користувач обирає у зрізі сегмент (наприклад, Loyal / Champions), після чого всі візуалізації на обох сторінках перебудовуються синхронно: KPI-картки показують показники цього сегмента, кругова діаграма відображає його стовідсоткову частку, географічна карта – регіональний розподіл, а матриця – детальні поведінкові показники. Це дозволяє всебічно оцінити обрану групу клієнтів. Приклад синхронізованої фільтрації за сегментом Loyal / Champions на обох сторінках дашборду наведено на рис. 3.7. та рис. 3.8.


Рисунок 3.7 – Синхронізована фільтрація за сегментом Loyal / Champions: сторінка «Огляд»

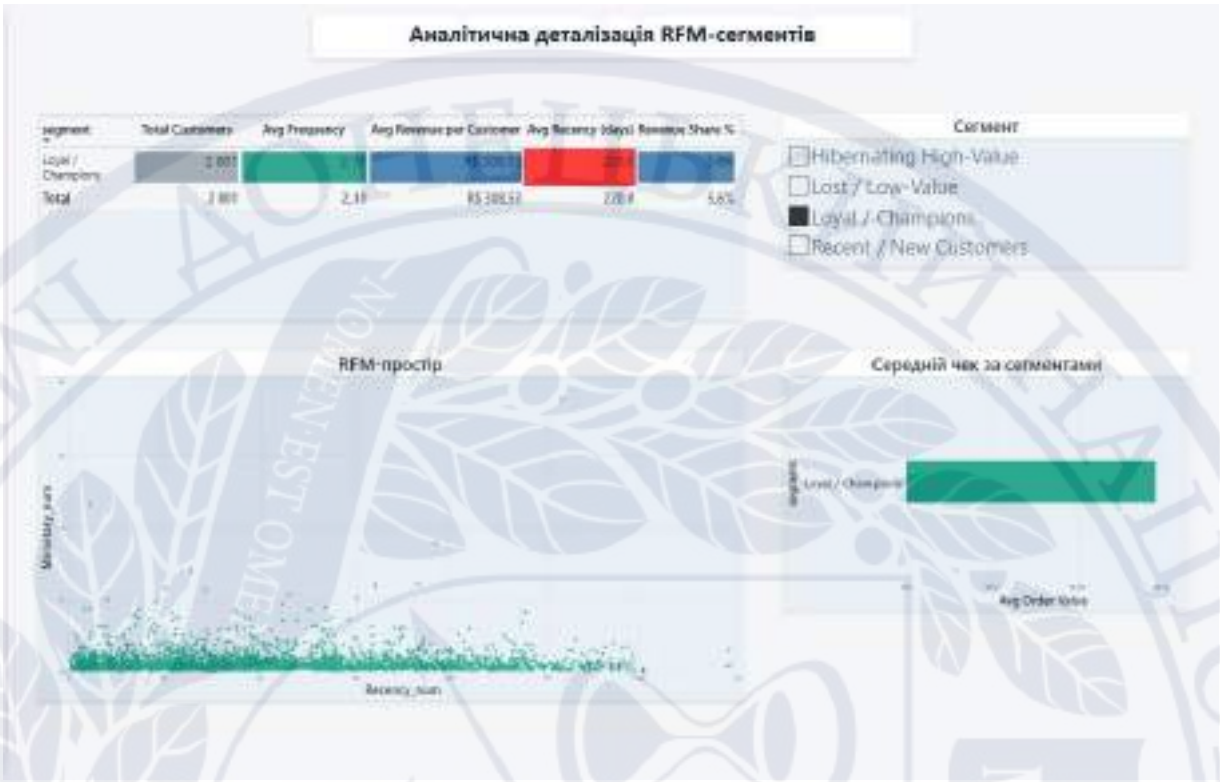


Рисунок 3.8 – Синхронізована фільтрація за сегментом Loyal / Champions: сторінка «Аналітика RFM»

### 3.11. Аналітична інтерпретація результатів та практична цінність системи

Розроблена аналітична система дозволяє сформулювати низку практично значущих висновків щодо клієнтської бази досліджуваного e-commerce маркетплейсу. Ключовим висновком є виявлення критичної концентрації доходу: сегмент Hibernating High-Value, охоплюючи лише 34,85 % клієнтів, генерує 61,85 % сумарного доходу, однак ці клієнти не здійснювали покупок у середньому понад дев'ять місяців. Це означає, що основне джерело доходу підприємства перебуває у зоні високого ризику відтоку, і реактивація саме цього сегмента має бути стратегічним пріоритетом.

Другим важливим висновком є висока частка одноразових покупців: понад 96 % клієнтів здійснили лише одну покупку. Це характерно для маркетплейс-моделі

ринку, що розвивається, та вказує на нерозвиненість механізмів формування лояльності. Сегмент Loyal / Champions, що демонструє повторні покупки, становить лише 3,00 % бази, що підкреслює значний потенціал зростання за рахунок стимулювання повторних транзакцій.

На основі проведеної сегментації сформульовано адаптовані маркетингові рекомендації для кожної групи. Для сегмента Hibernating High-Value доцільні win-back кампанії з персоналізованими пропозиціями та нагадуваннями; для Recent / New Customers – програми конверсії в лояльних клієнтів через welcome-серії та рекомендації супутніх товарів; для Loyal / Champions – програми утримання та VIP-привілеї; для Lost / Low-Value – мінімізація маркетингових витрат або виключення з активних кампаній.

Практична цінність розробленої системи полягає у кількох аспектах. По-перше, система забезпечує перехід від інтуїтивного до даних-керованого підходу в сегментації клієнтів, що підвищує обґрунтованість маркетингових рішень. По-друге, реалізація на основі відкритих та безкоштовних інструментів (R та Power BI Desktop) робить систему доступною для малого та середнього бізнесу. По-третє, модульна архітектура дозволяє адаптувати систему під дані інших підприємств без істотних модифікацій – достатньо замінити вхідний датасет, зберігши структуру RFM-показників. По-четверте, інтерактивний дашборд знижує бар'єр входу для нетехнічних користувачів, дозволяючи їм самостійно проводити аналітичні дослідження.

Таким чином, розроблена аналітична система сегментації клієнтів є завершеним практичним рішенням, що поєднує статистичну обґрунтованість методів машинного навчання з доступністю інтерактивної бізнес-аналітики, та може бути використана для підвищення ефективності маркетингової діяльності e-commerce підприємств.

**Висновки до розділу 3**

У третьому розділі розроблено та реалізовано аналітичну систему сегментації клієнтів, що інтегрує статистичну обробку даних мовою R з інтерактивною візуалізацією засобами Power BI. Спроєктовано архітектуру системи на основі принципу слабкої зв'язаності компонентів, що забезпечує її гнучкість, відтворюваність та можливість адаптації.

Реалізовано програмні модулі завантаження та інтеграції даних, формування RFM-показників, кластеризації методом K-means та експорту результатів, наведено ключові фрагменти коду.

Розроблено семантичну модель даних у Power BI, що включає обчислювані стовпці для коректного перетворення числових показників та DAX-метрик для обчислення ключових аналітичних показників.

Описано механізми інтерактивності дашборду та типові сценарії його використання. Проведено аналітичну інтерпретацію отриманих результатів, сформульовано адаптовані маркетингові рекомендації для кожного сегмента та обґрунтовано практичну цінність розробленої системи.

# ВИСНОВКИ

У кваліфікаційній (бакалаврській) роботі вирішено актуальне практичне завдання – розроблено аналітичну систему сегментації клієнтів електронної комерції на основі методів статистичного навчання із застосуванням інструментів бізнес-аналітики. Поставлену мету досягнуто, усі визначені завдання виконано, що дозволяє сформулювати такі висновки.

- Досліджено теоретичні основи сегментації клієнтів та методи статистичного навчання. Обґрунтовано доцільність використання поведінкової сегментації на основі транзакційних даних, а також алгоритму K-means у поєднанні з RFM-аналізом як методологічної основи системи.
- Виконано повний цикл підготовки даних: очищення, інтеграцію та формування RFM-показників для 93 349 клієнтів. Побудовано кластерну модель K-means, для якої оптимально визначено 4 сегменти, що забезпечили прийнятну якість кластеризації.
- Реалізовано програмну частину системи в R та аналітичний дашборд у Power BI з обчислюваними метриками та візуалізаціями, що забезпечує відтворюваність, наочність та інструментальну придатність рішення.
- Проінтерпретовано отримані сегменти та сформовано маркетингові рекомендації. Виявлено, що сегмент Hibernating High-Value генерує основну частку доходу (61,85 %), що визначає його як пріоритетний для реактивації. Окреслено перспективи подальшого розвитку системи, зокрема інтеграцію прогнозних моделей та автоматизацію оновлення сегментації.

Практична цінність роботи полягає у створенні завершеного рішення на основі відкритих інструментів (R та Power BI Desktop), доступного та придатного до адаптації під дані інших підприємств. Перспективними напрямами подальших

досліджень є інтеграція прогнозних моделей переходу клієнтів між сегментами, врахування часової динаміки купівельної поведінки та автоматизація оновлення сегментаційної моделі в режимі, наближеному до реального часу.

# ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Smith W. R. Product differentiation and market segmentation as alternative marketing strategies. Journal of Marketing. 2021. Vol. 21, No. 1. P. 3–8. URL: https://journals.sagepub.com/doi/10.1177/002224295602100102 (Дата звернення: 09.05.2026.)

2. Kotler P., Keller K. L. Marketing Management. 15th ed. Harlow : Pearson Education, 2016. 832 p. URL: https://dspace.vnbrims.org/handle/123456789/5050 (Дата звернення: 10.05.2026.)

3. Buttle F., Maklan S. Customer Relationship Management: Concepts and Technologies. 4th ed. London : Routledge, 2019. 460 p. DOI: https://doi.org/10.4324/9781351016551 (Дата звернення: 10.05.2026.)

4. Wedel M., Kamakura W. A. Market Segmentation: Conceptual and Methodological Foundations. 2nd ed. Boston : Springer, 2000. 382 p. URL: https://link.springer.com/article/10.1057/palgrave.jt.5740007 (Дата звернення: 10.05.2026.)

5. Tsiptsis K., Chorianopoulos A. Data Mining Techniques in CRM: Inside Customer Segmentation. Chichester : John Wiley & Sons, 2010. 372 p. URL: https://onlinelibrary.wiley.com/doi/book/10.1002/9780470685815 Tsiptsis K., Chorianopoulos A. Data Mining Techniques in CRM: Inside Customer Segmentation. Chichester : John Wiley & Sons, 2010. 372 p. (Дата звернення: 11.05.2026.)

6. Verhoef P. C., Kooge E., Walk N. Creating Value with Big Data Analytics: Making Smarter Marketing Decisions. London : Routledge, 2016. 304 p. DOI: https://doi.org/10.4324/9781315734750 (Дата звернення: 11.05.2026.)

7. Provost F., Fawcett T. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. Sebastopol : O'Reilly Media, 2013. 414 p. URL: https://www.researchgate.net/publication/256438799_Data_Science_for_Business (Дата звернення: 12.05.2026.)

8. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd ed. Waltham : Morgan Kaufmann, 2011. 744 p. URL: https://www.scribd.com/presentation/806858532/Data-Mining-Concepts-and-Techniques-Han-Kamber-Pei (Дата звернення: 08.05.2026.)

9. Davenport T. H., Harris J. G. Competing on Analytics: The New Science of Winning. Boston : Harvard Business School Press, 2006. 240 p. URL: https://www.researchgate.net/publication/7327312_Competing_on_Analytics (Дата звернення: 09.05.2026.)

10. Гафіяк А. М. IT-технології та бізнес-аналітика. *Економіка і суспільство*. 2018. Вип. 15. С.933–937 .URL: https://economyandsociety.in.ua/journals/15_ukr/143.pdf (Дата звернення: 10.05.2026.)

11. Sherman R. Business Intelligence Guidebook: From Data Integration to Analytics. Waltham : Morgan Kaufmann, 2014. 550 p. URL: https://surl.li/jgkmfx (Дата звернення: 10.05.2026.)

12. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning with Applications in R. 2nd ed. New York : Springer, 2021. 607 p. URL: https://www.statlearning.com/ (Дата звернення: 11.05.2026.)

13. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York : Springer, 2009. 745 p. URL: https://surl.li/nvqfsb (Дата звернення: 11.05.2026.)

14. Jain A. K. Data clustering: 50 years beyond K-means. Pattern Recognition Letters. 2010. Vol. 31, No. 8. P. 651–666. DOI: https://doi.org/10.1016/J.PATREC.2009.09.011 (Дата звернення: 10.05.2026.)

15. Xu R., Wunsch D. Survey of clustering algorithms. IEEE Transactions on Neural Networks. 2005. Vol. 16, No. 3. P. 645–678. DOI: https://doi.org/10.1109/TNN.2005.845141 (Дата звернення: 10.05.2026.)

16. Alves Gomes M., Meisen T. An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. *Analytics*. 2023. Vol. 2, No. 4. P. 809–

823. (Open Access, MDPI) URL: https://www.mdpi.com/2813-2203/2/4/42?utm_source=researchgate.net&utm_medium=article (Дата звернення: 12.05.2026.)

17. Lloyd S. P. Least squares quantization in PCM. IEEE Transactions on Information Theory. 1982. Vol. 28, No. 2. P. 129–137. URL: https://scispace.com/papers/least-squares-quantization-in-pcm-56gjks6m10 (Дата звернення: 13.05.2026.)

18. Rousseeuw P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987. Vol. 20. P. 53–65. DOI: https://doi.org/10.1016/0377-0427(87)90125-7 (Дата звернення: 10.05.2026.)

19. Tibshirani R., Walther G., Hastie T. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B. 2001. Vol. 63, No. 2. P. 411–423. DOI: https://doi.org/10.1111/1467-9868.00293 (Дата звернення: 10.05.2026.)

20. Tan P.-N., Steinbach M., Karpatne A., Kumar V. Introduction to Data Mining. 2nd ed. New York : Pearson, 2018. 864 p. URL: https://www.scribd.com/document/373063649/80d7effb4c3a4e1514eff6139a37a1028f1b (Дата звернення: 11.05.2026.)

21. Wong E., Wei Y. Customer Online Shopping Behaviour Data Analysis Using ML-Based Classification for Effective Segmentation. *Sensors*. 2023. Vol. 23, No. 6. 3180. URL: https://www.mdpi.com/1424-8220/23/6/3180 (Дата звернення: 12.05.2026.)

22. Bult J. R., Wansbeek T. Optimal Selection for Direct Mail. Marketing Science. 1995. Vol. 14, No. 4. P. 378–394. URL: https://econpapers.repec.org/article/inmormksc/v_3a14_3ay_3a1995_3ai_3a4_3ap_3a378-394.htm (Дата звернення: 10.05.2026.)

23. Wei J.-T., Lin S.-Y., Wu H.-H. A review of the application of RFM model. African Journal of Business Management. 2010. Vol. 4, No. 19. P. 4199–4206. URL:

https://www.researchgate.net/publication/228399859_A_review_of_the_application_of_RFM_model (Дата звернення: 09.05.2026.)

24. Tabianan K., Velu S., Ravi V. K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. *Sustainability*. 2022. Vol. 14, No. 12. 7243. DOI: https://doi.org/10.3390/su14127243 (Дата звернення: 13.05.2026.)

25. Christy A. J., Umamakeswari A., Priyatharsini L., Neyaa A. RFM ranking – An effective approach to customer segmentation. Journal of King Saud University – Computer and Information Sciences. 2021. Vol. 33, No. 10. P. 1251–1257. DOI: https://doi.org/10.1016/j.jksuci.2018.09.004 (Дата звернення: 10.05.2026.)

26. What is Power BI? Microsoft Learn. Дата звернення: 21.05.2026. URL: https://learn.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview (Дата звернення: 10.05.2026.)

27. Russo M., Ferrari A. The Definitive Guide to DAX. 2nd ed. Redmond : Microsoft Press, 2019. 768 p. URL: https://ptgmedia.pearsoncmg.com/images/9780735698352/samplepages/9780735698352.pdf (Дата звернення: 11.05.2026.)

28. Назаренко М. С., Січко Т. В. Аналіз клієнтських даних із застосуванням інструментів бізнес-аналітики. Вісник Донецького національного університету імені Василя Стуса. Вінниця. Випуск 18. Том 1. 2026. URL: https://jvestnik-sss.donnu.edu.ua/issue/view/682 (Дата звернення: 10.05.2026.)

29. Marz N., Warren J. Big Data: Principles and Best Practices of Scalable Real-Time Data Systems. Shelter Island : Manning Publications, 2015. 328 p. URL: https://surl.li/lsjqpz (Дата звернення: 11.05.2026.)

30. Anitha P., Patil M. M. RFM model for customer purchase behavior using K-Means algorithm. Journal of King Saud University – Computer and Information Sciences. 2022. Vol. 34, No. 5. P. 1785–1792. DOI: https://doi.org/10.1016/j.jksuci.2019.12.011 (Дата звернення: 10.05.2026.)

31. Customer Segmentation: A Comparative Study Using Clustering Algorithms. arXiv, 2024. URL: https://arxiv.org/pdf/2402.04103 (Дата звернення: 10.05.2026.)

32. Davenport T. H. The AI Advantage: How to Put the Artificial Intelligence Revolution to Work. Cambridge: MIT Press, 2018. 248 p. DOI: https://doi.org/10.1080/15228053.2020.1756084 (Дата звернення: 10.05.2026.)

33. Chen H., Chiang R. H. L., Storey V. C. Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly. 2012. Vol. 36, No. 4. P. 1165–1188. DOI: https://doi.org/10.2307/41703503 (Дата звернення: 10.05.2026.)

34. Press G. Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. Forbes. 23.03.2016. URL: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/ (Дата звернення: 13.05.2026.)

35. Olist, Sionek A. Brazilian E-Commerce Public Dataset by Olist. Kaggle, 2018. URL: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce (Дата звернення: 11.05.2026.)

36. E-commerce market in Brazil – Statistics & Facts. Statista Research Department, 2023. URL: https://www.statista.com/topics/4697/e-commerce-in-brazil/ (Дата звернення: 10.05.2026.)

37. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL: https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf (Дата звернення: 11.05.2026.)

38. Wickham H., Grolemund G. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media, 2023. 522 p. URL: https://r4ds.hadley.nz/ (Дата звернення: 10.05.2026.)

39. Kassambara A. Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning. STHDA, 2017. 188 p. URL: https://www.datanovia.com/en/product/practical-guide-to-cluster-analysis-in-r/ (Дата звернення: 12.05.2026.)

40. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 3rd ed. New York : Springer, 2016. 260 p. URL: https://ggplot2-book.org/ (Дата звернення: 11.05.2026.)

41. Power BI report and dashboard creation documentation. URL: https://learn.microsoft.com/en-us/power-bi/create-reports/ (Дата звернення: 12.05.2026.)

42. Power Query M formula language. URL: https://learn.microsoft.com/en-us/powerquery-m/ (Дата звернення: 14.05.2026)

43. Data Analysis Expressions (DAX). URL: https://learn.microsoft.com/en-us/dax/ (Дата звернення: 14.05.2026)

# ДОДАТКИ

# ДОДАТОК А

Програмний код завантаження, очищення та інтеграції даних (мова R)

```
install.packages(c("tidyverse", "lubridate", "cluster", "factoextra",
"NbClust"))

library(tidyverse)
library(lubridate)
library(cluster)
library(factoextra)

setwd("C:/olist")
dir.create("output", showWarnings = FALSE)

set.seed(42)

orders   <- read_csv("olist_orders_dataset.csv",       show_col_types
= FALSE)
customers         <-          read_csv("olist_customers_dataset.csv",
show_col_types = FALSE)
items    <- read_csv("olist_order_items_dataset.csv",  show_col_types
= FALSE)
payments                                                          <-
read_csv("olist_order_payments_dataset.csv",show_col_types = FALSE)

cat(" ОБСЯГИ ВХІДНИХ ТАБЛИЦЬ \n")
cat("orders:   ", nrow(orders),    "\n")
cat("customers:",       nrow(customers),       "       (унікальних:",
n_distinct(customers$customer_unique_id), ")\n")
cat("items:    ", nrow(items),     "\n")
cat("payments: ", nrow(payments),  "\n\n")

cat(" ОЧИЩЕННЯ ДАНИХ \n")
cat("Розподіл статусів замовлень:\n")
print(table(orders$order_status))

orders_clean <- orders %>%
  filter(order_status == "delivered") %>%
  filter(!is.na(order_delivered_customer_date))
cat("\nПісля фільтрації delivered + наявність дати доставки:",
    nrow(orders_clean), "замовлень\n")
orders_clean <- orders_clean %>%
  mutate(order_purchase_timestamp                                   =
ymd_hms(order_purchase_timestamp))

order_totals <- items %>%
  group_by(order_id) %>%
  summarise(
```

```
    order_total_price   = sum(price),
    order_total_freight = sum(freight_value),
    n_items             = n(),
    .groups = "drop"
  ) %>%
  mutate(order_total = order_total_price + order_total_freight)

df <- orders_clean %>%
  inner_join(customers, by = "customer_id")

df <- df %>%
  inner_join(order_totals, by = "order_id")

cat("\nПідсумкова транзакційна таблиця:", nrow(df), "рядків,",
    n_distinct(df$customer_unique_id), "унікальних клієнтів\n\n")

ref_date <- max(df$order_purchase_timestamp) + days(1)
cat(" RFM-АНАЛІЗ \n")
cat("Референсна дата аналізу:", format(ref_date, "%Y-%m-%d"), "\n\n")

rfm <- df %>%
  group_by(customer_unique_id) %>%
  summarise(
    Recency                    =       as.numeric(difftime(ref_date,
max(order_purchase_timestamp), units = "days")),
    Frequency = n_distinct(order_id),
    Monetary  = sum(order_total),
    .groups   = "drop"
  )

cat("RFM-таблиця сформована:", nrow(rfm), "клієнтів\n")
cat("\nОписова статистика RFM:\n")
print(summary(rfm[, c("Recency", "Frequency", "Monetary")]))

skewness <- function(x) {
  m <- mean(x); s <- sd(x)
  mean((x - m)^3) / s^3
}
cat("\n АСИМЕТРІЯ РОЗПОДІЛІВ \n")
cat(sprintf("Recency  : raw = %.2f | log = %.2f\n",
            skewness(rfm$Recency),  skewness(log1p(rfm$Recency))))
cat(sprintf("Frequency: raw = %.2f | log = %.2f\n",
            skewness(rfm$Frequency),skewness(log1p(rfm$Frequency))))
cat(sprintf("Monetary : raw = %.2f | log = %.2f\n",
            skewness(rfm$Monetary), skewness(log1p(rfm$Monetary))))

cat("\n КОРЕЛЯЦІЙНА МАТРИЦЯ RFM \n")
print(round(cor(rfm[, c("Recency", "Frequency", "Monetary")]), 3))
```

```
p_dist <- rfm %>%
  pivot_longer(c(Recency, Frequency, Monetary), names_to = "metric",
values_to = "value") %>%
  ggplot(aes(value)) +
  geom_histogram(bins = 40, fill = "grey40", color = "white") +
  facet_wrap(~metric, scales = "free") +
  labs(title = "Розподіли RFM-показників", x = NULL, y = "Кількість
клієнтів") +
  theme_minimal(base_size = 12)
ggsave("output/fig_rfm_distributions.png", p_dist, width = 10, height
= 4, dpi = 150)

m_cap <- quantile(rfm$Monetary, 0.99)
cat(sprintf("\n99-й перцентиль Monetary: R$ %.2f (значення вище
обмежено)\n", m_cap))
rfm <- rfm %>% mutate(Monetary_capped = pmin(Monetary, m_cap))

rfm <- rfm %>%
  mutate(
    log_R = log1p(Recency),
    log_F = log1p(Frequency),
    log_M = log1p(Monetary_capped)
  )

X <- scale(rfm[, c("log_R", "log_F", "log_M")])

cat("\n ВИЗНАЧЕННЯ ОПТИМАЛЬНОГО K \n")

wcss <- sapply(2:7, function(k) {
  km <- kmeans(X, centers = k, nstart = 10, iter.max = 50)
  km$tot.withinss
})

set.seed(42)
sample_idx <- sample(seq_len(nrow(X)), 5000)
X_sample <- X[sample_idx, ]
sil <- sapply(2:7, function(k) {
  km <- kmeans(X_sample, centers = k, nstart = 10, iter.max = 50)
  ss <- silhouette(km$cluster, dist(X_sample))
  mean(ss[, 3])
})

k_eval <- data.frame(K = 2:7, WCSS = round(wcss, 1), Silhouette =
round(sil, 3))
cat("\nТаблиця оцінювання K:\n")
print(k_eval)
write_csv(k_eval, "output/k_selection.csv")
p_elbow <- ggplot(k_eval, aes(K, WCSS)) +
```

```
  geom_line(color = "steelblue") + geom_point(size = 3, color =
"steelblue") +
  labs(title = "Метод ліктя (Elbow method)", x = "Кількість кластерів
K", y = "WCSS") +
  theme_minimal(base_size = 12)
ggsave("output/fig_elbow.png", p_elbow, width = 6, height = 4, dpi =
150)

p_sil <- ggplot(k_eval, aes(K, Silhouette)) +
  geom_line(color = "darkorange") + geom_point(size = 3, color =
"darkorange") +
  labs(title = "Коефіцієнт силуету", x = "Кількість кластерів K", y =
"Середній силует") +
  theme_minimal(base_size = 12)
ggsave("output/fig_silhouette.png", p_sil, width = 6, height = 4, dpi
= 150)

cat("\n ПОБУДОВА МОДЕЛІ K-MEANS (K=4) \n")
set.seed(42)
km_final <- kmeans(X, centers = 4, nstart = 25, iter.max = 50)

rfm$cluster <- km_final$cluster
cat("WCSS (tot.withinss):", round(km_final$tot.withinss, 1), "\n")
cat("Between/Total SS:   ", round(km_final$betweenss / km_final$totss,
3), "\n\n")

cat("Розміри кластерів:\n")
print(table(rfm$cluster))

profile <- rfm %>%
  group_by(cluster) %>%
  summarise(
    n          = n(),
    R_mean     = round(mean(Recency), 1),
    F_mean     = round(mean(Frequency), 2),
    M_mean     = round(mean(Monetary), 2),
    M_sum      = round(sum(Monetary), 2),
    .groups    = "drop"
  ) %>%
  mutate(
    share_pct     = round(n / sum(n) * 100, 2),
    rev_share_pct = round(M_sum / sum(M_sum) * 100, 2)
  )
cat("\n=== ПРОФІЛІ СЕГМЕНТІВ ===\n")
print(as.data.frame(profile))
write_csv(profile, "output/cluster_profile.csv")

seg_names <- profile %>%
  mutate(segment = case_when(
```

```
    F_mean > 1.5                              ~ "Loyal / Champions",
    R_mean < 100                              ~ "Recent / New Customers",
    M_mean > median(profile$M_mean)           ~ "Hibernating High-Value",
    TRUE                                      ~ "Lost / Low-Value"
  )) %>%
  select(cluster, segment)

rfm <- rfm %>% left_join(seg_names, by = "cluster")
cat("\n РОЗПОДІЛ ПО СЕГМЕНТАХ \n")
print(table(rfm$segment))

p_clusters <- fviz_cluster(km_final, data = X,
                           geom = "point", alpha = 0.3,
                           palette    =    "Set2",    ggtheme    =
theme_minimal()) +
  labs(title = "Кластери клієнтів (простір головних компонент)")
ggsave("output/fig_clusters_pca.png", p_clusters, width = 7, height =
5, dpi = 150)

p_scatter <- ggplot(rfm, aes(Recency, Monetary, color = segment)) +
  geom_point(alpha = 0.3, size = 0.8) +
  scale_y_log10() +
  labs(title = "Сегменти клієнтів: Recency vs Monetary",
       x = "Recency (днів)", y = "Monetary (R$, лог-шкала)", color =
"Сегмент") +
  theme_minimal(base_size = 12)
ggsave("output/fig_scatter_segments.png", p_scatter, width = 8, height
= 5, dpi = 150)

geo <- df %>%
  group_by(customer_unique_id) %>%
  summarise(customer_state = first(customer_state),
            customer_city  = first(customer_city), .groups = "drop")

powerbi_export <- rfm %>%
  left_join(geo, by = "customer_unique_id") %>%
  select(customer_unique_id, Recency, Frequency, Monetary,
         cluster, segment, customer_state, customer_city)

write_csv(powerbi_export, "output/olist_segments_powerbi.csv")
cat("\n>>>       Файл       для       Power       BI       збережено:
output/olist_segments_powerbi.csv\n")
cat(">>> Рядків:", nrow(powerbi_export), "\n")
```