КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Економічний факультет
Кафедра економічної кібернетики

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

«Аналітика клієнтів за допомогою сегментації на основі даних компанії Airbnb»

студентки 4 курсу
спеціальності 051 «Економіка»
ОПП «Економічна кібернетика»
денної форми навчання
Кожановської Ольги Сергіївни

Науковий керівник:
д.е.н., професор
Чорноус Галина Олександрівна

Засвідчую, що у цій дипломній роботі немає запозичень із праць інших авторів без відповідних посилань
Студент_____
(підпис)

Роботу допущено до захисту перед ЕК
рішенням кафедри економічної кібернетики
від 12 червня 2023 р., протокол № 17
Завідувач кафедри:
доктор економічних наук, професор
Ляшенко Олена Ігорівна

_____
(підпис)

КИЇВ – 2023

РЕФЕРАТ

Кваліфікаційна робота бакалавра містить: 66 ст., 16 рис., 3 табл., 43 джерел, додатки.

Ключові слова: економіка спільного споживання, Airbnb, сегментація користувачів, алгоритми кластеризації, кластерний аналіз.

Об’єкт дослідження: процеси прийняття рішень в сфері оренди житла та їх аналітична підтримка.

Мета дослідження: аналіз клієнтської бази компанії Airbnb за допомогою використання алгоритмів неконтрольованого навчання.

Методи дослідження: K-Means, DBSCAN, BIRCH.

Наукова новизна, теоретична значимість дослідження: запропоновано авторський комплексний підхід до сегментації клієнтської бази із застосуванням алгоритмів машинного навчання для побудови ефективних маркетингових стратегій. Побудовано декілька моделей кластеризації, розглянуто їх різницю та проведено порівняльний аналіз.

Теоретична значимість дослідження полягає у розширенні бази існуючих досліджень в області сегментації клієнтів та їхнього аналізу з урахуванням особливостей компанії Airbnb. Це дозволить розкрити специфіку галузі гостинності та подорожей, виявити унікальні риси та особливості клієнтів, які використовують Airbnb як свою основну платформу для бронювання помешкань. Дослідження такого роду в контексті Airbnb ще недостатньо представлені в наукових джерелах, тому ця робота вносить нові знання і доповнює наявну літературу, сприяючи розвитку теорії сегментації клієнтів в сфері гостинності.

Практична цінність: результати дослідження є підґрунтям цінної інформацію про поведінку та вимоги різноманітних сегментів клієнтів, що дозволить розроблювати стратегії покращення маркетингової тактики, персоналізувати пропозиції та підвищити рівень задоволеності клієнтів.

ЗМІСТ

ВСТУП

Актуальність теми дослідження. В останні роки цифрові платформи набули все більшого значення в секторі оренди житла та змінили моделі споживання в інших сферах повсякденного життя (наприклад, послуги доставки їжі). Сьогодні такі платформи служать постачальниками або посередниками для обміну різними ресурсами. Зокрема, на основі цифрових платформ бізнес-модель економіки спільного споживання набула значного поширення в туристичному секторі і суміжних галузях господарства, культури, науки та освіти.

Індустрія економіки спільного споживання трансформувала ринок короткострокової оренди та сприяла зростанню популярності послуг різних компаній, як от Airbnb, що дало можливість розробити нові методи та підходи для аналізу поведінки та потреб клієнтів. Крім того, зростання впливу онлайн-платформи Airbnb у багаторівневій економіці призвело до збільшення пропозиції житла та конкуренції на ринку туристичних послуг, що дозволило споживачам мати вибір з більш широкого спектру варіантів проживання і знизило витрати на туристичні послуги.

Поява Airbnb також привела до зростання конкуренції на ринку туристичних послуг. Традиційні готелі та інші види проживання змушені адаптуватись до нових реалій та конкурувати з нещодавно утвореними платформами економіки спільного споживання. Це призвело до поліпшення якості обслуговування, зниження цін та появи нових інновацій у готельній галузі, спрямованих на привернення та утримання клієнтів.

Туристична галузь, включаючи послуги оренди житла, продовжує розвиватися і адаптуватися до змінних умов, враховуючи нові тренди, попит та потреби споживачів. Аналітика клієнтів та сегментація є важливим інструментом для розуміння та задоволення цих потреб, що робить тему дипломної роботи актуальною на сьогодні.

Дослідження сегментації на основі даних Airbnb є особливо актуальним в контексті зростання конкуренції в галузі гостьового розміщення. Зараз на ринку існує велика кількість подібних сервісів, тому для компанії, яка пропонує послуги в цій сфері, важливо мати ефективну маркетингову стратегію, яка дозволить привернути та утримувати клієнтів.

Розробки у сфері штучного інтелекту, машинного навчання та обробки великих обсягів даних надають нові можливості для збору, аналізу та використання даних клієнтів. Використання таких технологій у поєднанні з даними компанії Airbnb дозволяє отримати більш глибокі та точні патерни щодо поведінки та вподобань клієнтів. Саме тому, в умовах, коли технологічний прогрес продовжує стрімко розвиватися, дослідження з аналітики клієнтів має досить велику актуальність.

Беручи до уваги недостатню кількість вже наявних досліджень стосовно сегментації клієнтської бази саме компанії Airbnb, дана дипломна робота спрямована на розширення цієї тематики та впровадженню нових підходів до сегментації за допомогою впровадження методів машинного навчання.

Зокрема, не можна не згадати, що від початку війни суттєво змінився попит та фокус на туристичні послуги, включаючи бронювання помешкань через платформу Airbnb, у першу чергу, в Україні. Аналітика клієнтів за допомогою сегментації має на меті зрозуміти, як саме змінився попит на різні види помешкань, які географічні зони показують знижений або зростаючий інтерес, які клієнтські сегменти залишилися активними тощо. Це може допомогти визначити нові стратегії маркетингу та пристосуватися до нових умов.

Об’єктом дослідження є процеси прийняття рішень в сфері оренди житла та їх аналітична підтримка.

Предметом дослідження, в свою чергу, є моделі та методи неконтрольованого машинного навчання та аналітичні методи, що підтримують процеси прийняття рішень у сфері оренди житла.

Метою дипломної роботи є проведення аналізу клієнтів компанії Airbnb за допомогою сегментації на основі наявних даних, включаючи моделі неконтрольованого машинного навчання. Зокрема, основні завдання дослідження включають:

- опис місця компанії Airbnb в системі глобального ринку туристичних послуг;
- виявлення основних сегментів клієнтів Airbnb на основі демографічних, поведінкових та географічних характеристик;
- аналіз бази даних, на якій базується дослідження;
- пошук поведінкових та споживчих патернів різних сегментів клієнтів;
- оцінка ефективності реалізованих моделей кластеризації;
- розробка рекомендацій та стратегій для поліпшення взаємодії з різними сегментами клієнтів та збільшення їхньої задоволеності.

Методи дослідження в даній роботі включають моделі кластеризації на основі неконтрольованого машинного навчання, серед яких: K-Means, DBSCAN, BIRCH. Практична реалізація виконана на мові програмування Python на основі даних, які представлені на сайті «Inside Airbnb».

Наукова новизна полягає в систематизації теоретичних напрацювань та практики моделей кластеризації, виявленні основних тенденцій підходу до сегментації клієнтів Airbnb. Запропоновано використовувати алгоритм BIRCH, який раніше не використовувався в подібних дослідженнях сегментації для обраної компанії. Представлено рейтинг моделей машинного навчання та оцінено їх ефективність.

Практичне значення. Дана дипломна робота може бути корисною для проектування аналізу компаній, які пропонують послуги у сфері гостьового розміщення, а також для дослідників, які займаються вивченням методів сегментації на основі даних та їх подальшим застосуванням у різних дослідженнях.

Структура дослідження. Робота структурно складається зі вступу, трьох розділів, висновків і списку використаних джерел зі 43 найменувань. Повний обсяг роботи становить 66 сторінок, містить 3 таблиці, 16 рисунків та додаток.

РОЗДІЛ 1. ОГЛЯД ТА ВИЗНАЧЕННЯ МІСЦЯ AIRBNB В ЕКОНОМІЦІ СПІЛЬНОГО СПОЖИВАННЯ

1.1. Економіка спільного споживання

З поширенням цифрових технологій та діджиталізацією економіки за останні 15 років активно почали з’являтись різноманітні бізнес-моделі, які дозволяють компаніям ефективніше працювати, зменшувати свої витрати на проведення різних операцій та збільшувати свої прибутки. Таким чином цифрова трансформація економіки стала початком створення нового явища – економіки спільного споживання, або шерингової економіки. Це економічне явище являє собою соціоекономічну систему, що заснована на шерингу, тобто спільному користуванні людськими та фізичними ресурсами [1].

Економіка спільного споживання є досить стійкою економічною системою, що побудована навколо спільного використання приватних активів. Ця система здебільшого покладається на інформаційні технології (P2P), для надання можливості окремим особам та іншим прибутковим і некомерційним організаціям обмінюватися надлишковими ресурсами товарів, знань і послуг. Враховуючи той факт, що надійність і безпека є вирішальними факторами, коли справа доходить до спільного використання приватних активів, вартість цих активів збільшується лише завдяки обміну інформацією про них [2].

Основною характеристикою шерингової економіки є можливість продавати товари та послуги без посередників, що значно полегшує процес та дозволяє більш ефективно використовувати ресурси. Різноманітні IT-розробки та застосування Big Data, що дозволяють розміщувати інформацію про послуги та товари на децентралізованих порталах і платформах, стали важливою частиною моделі економіки спільного споживання. Особлива перевага моделі спільного споживання для малих і середніх підприємств полягає в тому, що вони можуть легко вийти на світовий ринок і забезпечити швидкий розвиток. Зокрема, у Європі ринок зростає на 35% щорічно і наразі становить 335 мільярдів доларів [3].

Напрям економіки спільного споживання є досить новим та активно розвивається, тому на сьогодні не існує однозначного визначення поняття шерингової економіки. Попри те, що суперечки щодо того, які саме напрями охоплює економіка спільного споживання досі тривають, вже зараз можна зробити висновок, що ця ринкова модель має експоненціальний ріст як з точки зору компаній, діяльність та бізнес-модель яких базується на ідеї економіки спільного споживання, так і з точки зору цінності споживачів. Відповідно, кількість досліджень, присвячених економіці спільного споживання зростає в геометричній прогресії.

Для розуміння поточного стану різних досліджень та визначення майбутніх напрямів розвитку економіки спільного споживання науковці розробили систематизовані комплексні огляди. Наприклад, М. Хоссейн в своїй праці «Sharing economy: A comprehensive literature review» [4] провів аналіз 129 наукових робіт та розробив механізм пояснення явища економіки спільного споживання, а також формулює майбутні напрями досліджень у цій галузі. Даний механізм визначає, що економіка спільного споживання – галузь, що є предметом вивчення багатьох дисциплін: економіки, яка цікавиться моделями бізнесу; юриспруденції, що аналізує явище з точки зору регулювання політики; соціології, яка досліджує поведінку та мотивацію споживачів тощо.

Р. Белк [5] також зазначає, що економіка спільного споживання, яка включає економічні і соціальні елементи, означає саме дію спільного споживання, а не купівлю речей чи відсутність власності. Прайс [6] зазначив, що спільне споживання це «найбільш універсальна форма економічної поведінки, пов’язана з людиною», що Р. Белк пояснює як віддачу будь-чого іншим споживачам для використання або отримання будь-якої речі від споживачів з метою власного використання. Також Р. Белк наголошує, що якщо під час передачі речі або послуги в спільне використання сплачується будь-яка плата чи компенсація, таке явище слід класифікувати як «спільне споживання» та розглядати виключно як концепцію економіки спільного споживання [7].

Економіка спільної діяльності, спільне споживання, економіка доступу, економіка платформи, економіка спільноти, шерингова економіка - кожен із цих термінів відображає основні аспекти цього багатогранного явища. Однак одне з найточніших визначень було сформульоване у 2017 році Р. Муньосом та Д. Коеном, які визначають економіку спільного використання як «соціально-економічну систему, яка забезпечує обмін товарами та послугами між окремими особами та організаціями, призначену для підвищення ефективності та оптимізацію недостатньо використовуваних ресурсів у суспільстві» [8].

Шерингова економіка є результатом постійного розвитку та інновацій у просторі цифрових технологій. Користувачі Інтернету перейшли від пасивних одержувачів інформації до активних виробників і споживачів інформації (прикладом можна назвати користування такими ресурсами, як Wikipedia, Facebook, Twitter, Tripadvisor, Zomato, Airbnb тощо), а останнім часом активного поширення набув обмін недостатньо використовуваними активами, такими як житлова площа, машини, час і навички.

Оскільки на сьогодні мобільні технології дозволяють легко отримати доступ до Інтернету, все більше і більше бізнес-діяльності здійснюється онлайн за допомогою цифрових платформ. Це заохочує багатьох звичайних людей безкоштовно приєднуватися до таких цифрових платформ і ділитися потенційно недостатньо використаними ресурсами з усіма користувачами платформи. Airbnb і Uber є лише двома прикладами альтернативних ринків, де витрати на трансакцію нижчі, ніж у традиційних організацій у галузі, а бар’єри для входу нижчі. Наприклад, Airbnb – це економічна платформа спільного використання, яка підтримує короткострокову оренду житла між власниками та орендарями, тоді як Uber – це економічна платформа спільного використання, яка полегшує подорожі на приватних автомобілях для окремих осіб і груп.

Однак, концепція економіки спільного споживання є досить різносторонньою з тої точки зору, що існують різні практики обміну, в залежності від яких визначається бізнес-модель компанії. Наприклад, у випадку

Uber, водії є радше виробниками послуги, ніж споживачами, навіть якщо вони використовують власне авто для поїздок. Це відрізняється від таких платформ спільного використання, як BlaBlaCar, де водій проїхав би маршрут у будь-якому випадку, незалежно від того, чи подорожують з ним користувачі платформи, таким чином одночасно створюючи поїздку та отримуючи кошти в подорожі.

1.2. Економіка спільного споживання та індустрія туризму

Вплив економіки спільного споживання є надзвичайно важливим у зростанні рівня розвитку різних галузей економіки, забезпечуючи потужне підґрунтя для створення нових інноваційних систем та технологій. Наприклад, нові здобутки можуть слугувати появі нових технологій, за допомогою яких можна забезпечити оптимальне використання ресурсів, скорочення відходів і підвищення забезпечення сталого розвитку. Також збільшення кількості P2P-платформ стимулює розвиток малих і середніх підприємств, що збільшує кількість зайнятого населення та підвищує економічну активність у тому чи іншому регіоні.

Найбільший вплив впровадження концепції спільного споживання можна спостерігати в галузях, що відносяться до третинного сектору економіки. Серед них варто виокремити галузь туристичних та готельних послуг, на основі якої проводиться аналітичне дослідження даної кваліфікаційної роботи.

Зокрема, відповідно до щорічного звіту WTTC (World Travel & Tourism Council) у 2022 році сектор подорожей і туризму склав близько 7,6% світового ВВП, при цьому відбулось зростання на 22% порівняно з попереднім роком. Також у 2022 році в туристичній галузі було створено понад 22 мільйони нових робочих місць, що загалом позитивно впливає на економічний розвиток певних країн чи регіонів. Витрати місцевих відвідувачів в туристичному секторі в 2022 році збільшились на 20,4% порівняно з 2021 роком, а витрати міжнародних відвідувачів зросли на 81,9%. До пандемії Covid-19 дана галузь створювала 1 з 5 робочих місць в світі протягом 2014-2019 років та генерувала 10,4% глобального

ВВП, що складає понад 10 трильйонів доларів США у 2019 році [9]. Незважаючи на зростання, обсяг туристичного ринку в усьому світі залишився нижчим за рівень до пандемії, склавши понад 1,9 трильйонів доларів США у 2022 році. Відповідно до прогнозу, можна очікувати, що ця сума зросте майже до 2,29 трильйона доларів США у 2023 році, перевищивши пік, який спостерігався в 2019 році [10].

Варто також згадати, що за новими даними UNWTO (World Tourism Organization) у 2022 році у всіх регіонах світу спостерігається значне зростання кількості міжнародних туристів, біженців та переміщених осіб, понад 900 мільйонів осіб здійснили міжнародні подорожі. Це все ще 63% від рівня до пандемії коронавірусу, але вдвічі більше, ніж у 2021 році. На Близькому Сході спостерігалося найвище відносне зростання, оскільки кількість прибуттів зросла до 83% від рівня до пандемії. Європа прийняла близько 585 мільйонів відвідувачів у 2022 році, що становить майже 80% від рівня до пандемії. Африка та Америка відновили близько 65% відвідувань до пандемії, тоді як Азіатсько-Тихоокеанський регіон досяг лише 23% через посилені запобіжні обмеження, які лише нещодавно почали послаблювати [11].

Якщо раніше всі основні послуги в регіоні надавали туристичні агентства, готелі, пансіонати, транспортні авіалінії та залізниці, то на сьогодні, не зважаючи на значне підвищення рівня туристичної галузі в світі, дані послуги вже не грають ключову роль на ринку. Оскільки розвиток економіки спільного споживання та туристичного сектору (зокрема, готельних послуг) почали активно розвиватись практично одночасно, поява P2P моделі істотно змінила напрямок розвитку та спосіб ведення туристичних, готельних послуг та послуг оренди. Разом з цим значна кількість інших послуг та сегментів безпосередньо пов’язані з індустрією гостинності. Наприклад, вагомих змін зазнає транспортна галузь у всіх наданих послугах (авіаперевезення, залізнична мережа, оренда автомобілів тощо), оскільки з’являються нові інноваційні рішення, що потребують впровадження нових технологій та систем.

Не можна не згадати про ще один ключовий фактор успіху туризму – збільшення кількості онлайн-платформ, різноманітних сайтів, маркетплейсів, розвиток онлайн-ринку електронної комерції, популяризація соціальних мереж та в цілому цифровізація різних сфер повсякденного життя. Такі онлайн-платформи дозволяють отримати доступ до зростаючої кількості інформації, створюючи більше можливостей для розширення напрямків розвитку та обізнаності населення про перспективи туристичної галузі. Таким чином туристичний та житловий сектори залучають цифрові інструменти для залучення нових споживачів, використовуючи соціальні мережі, туристичні спільноти, блоги та інші платформи.

Інформаційно-комунікаційні технології відіграють фундаментальну роль в електронному туризмі, оскільки компанії можуть підвищити свою ефективність та результативність, запровадивши оцифрування процесів та ланцюжків створення вартості в індустрії туризму, подорожей, гостинності та громадського харчування [12].

Отже, інформаційні технології дозволяють туристичній індустрії збирати та аналізувати величезні обсяги даних про поведінку клієнтів, уподобання та демографічні показники, які можна використовувати для сегментації клієнтів на різні групи, в залежності від певних характеристик, інтересів та поведінки.

1.3. Airbnb як приклад цифрової платформи економіки спільного споживання

В останні десятиліття послуги на цифрових платформах набувають все більшого значення в туристичному секторі економіки. Концепція обміну або спільного використання житла набула значного поширення у глобальному масштабі та згодом перетворилась на успішні бізнес-моделі, які мають вагомий вплив на традиційні провайдери послуг і місцеві ринки праці.

Яскравим прикладом впровадження такої моделі є поява на ринку компанії Airbnb, оскільки саме ця компанія є одним з найбільш вдалих прикладів розуміння та адаптації до змін у споживчому погляді на подорожування та

оренду нерухомості. Надаючи можливість споживачам бронювати житло в приватних будинках і квартирах, дана компанія зробила революцію в традиційній індустрії гостинності та значно розширила вибір і знизила витрати на житло. Саме це зробило Airbnb однією з найуспішніших технологічних компаній сьогодення та прикладом онлайн-платформи економіки спільного використання, яка користується широкою популярністю і значно впливає на спосіб бронювання подорожей і житла.

Компанія Airbnb була створена 2008 року, і з тих пір лише набирає оберти, ставши конкурентоспроможною та популярною альтернативою традиційним готелям, хостелам та іншим видам орендованого житла [13].

Згідно з статистикою, в 2023 році на платформі налічується понад 4 мільйони хостів Airbnb у всьому світі та близько 6 мільйонів активних списків на платформі. Оголошення Airbnb охоплюють щонайменше 100 тисяч міст та займають понад 20% ринку оренди нерухомості. Також Airbnb має понад 150 мільйонів користувачів у всьому світі, які забронювали понад 1 мільярд місць станом на 2023 рік. Зокрема, щосекунди на платформі реєструються 6 гостей, середня вартість їх оренди становить близько 163 долари за ніч, тривалість оренди – 4,3 ночі. На бронювання проживання в додатку Airbnb користувачам потрібно в середньому 11 хвилин і 31 секунду. Середній заробіток хоста становить понад 13 800 доларів на рік [14].

Компанія також співпрацює з 8700 сторонніми партнерами, а також понад 400 тисяч компаній користуються послугами Airbnb для організації подорожей своїх співробітників. Компанія також має понад 400 угод з місцевими та національними урядами щодо автоматизації збору туристичних податків та в рамках цих угод зібрала понад 2 мільярди доларів США податків, що пов’язані з туризмом [15].

Варто також згадати, що станом на травень 2023 року ринкова капіталізація Airbnb оцінюється в 75,55 мільярда доларів, порівняно зі 113 мільярдами доларів у 2021 році, що дозволило стати Airbnb 192 найдорожчою компанією у світі [16].

Через 12 років існування у 2020 році компанія вийшла на біржовий ринок, зокрема, у лютому 2023 року ціна акцій Airbnb становила 129 доларів США, тоді як у лютому 2021 року історичний максимум становив 212,68 доларів США [17]. У 2020 році Airbnb оцінив своє IPO в 68 доларів за акцію, але торги відкрились за ціною 146 доларів за акцію. Також на початку 2020 року кількість співробітників Airbnb досягла 7500 осіб, але понад 1900 співробітників були звільнені через вплив пандемії коронавірусу на бізнес [18].

На зображенні нижче можна побачити статистику квартального доходу Airbnb з 2014 по 2023 рік (мм дол. США). Зокрема, доходи Airbnb підвищились на 40% у 2022 році, тобто спостерігається зростання після зниження доходів на 31% у 2020 році через початок епідемії COVID-19.

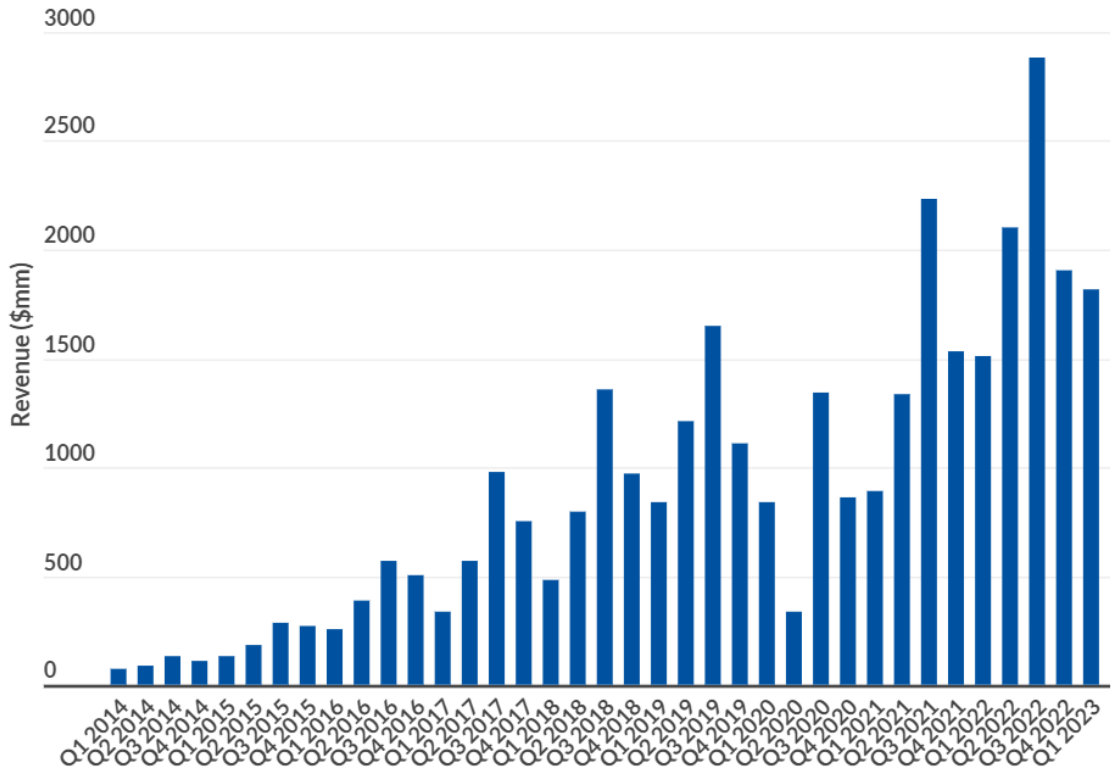

Рис. 1.1. Квартальний дохід Airbnb за 2014-2023 роки

Джерело: [19]

Якщо порівнювати частку Airbnb та ринку послуг з оренди в 2021 році, то компанія займає 2 місце, охоплюючи 22,9% всіх споживачів, в той час як Booking надає 34,82% всіх послуг ринку. Такий розрив може пояснюватись тим, що Airbnb відносно нова на ринку та має досить унікальну стратегію розвитку.

Також вплив має схильність клієнтів віддавати перевагу відомим брендам, вважаючи їх надійнішими та безпечнішими.

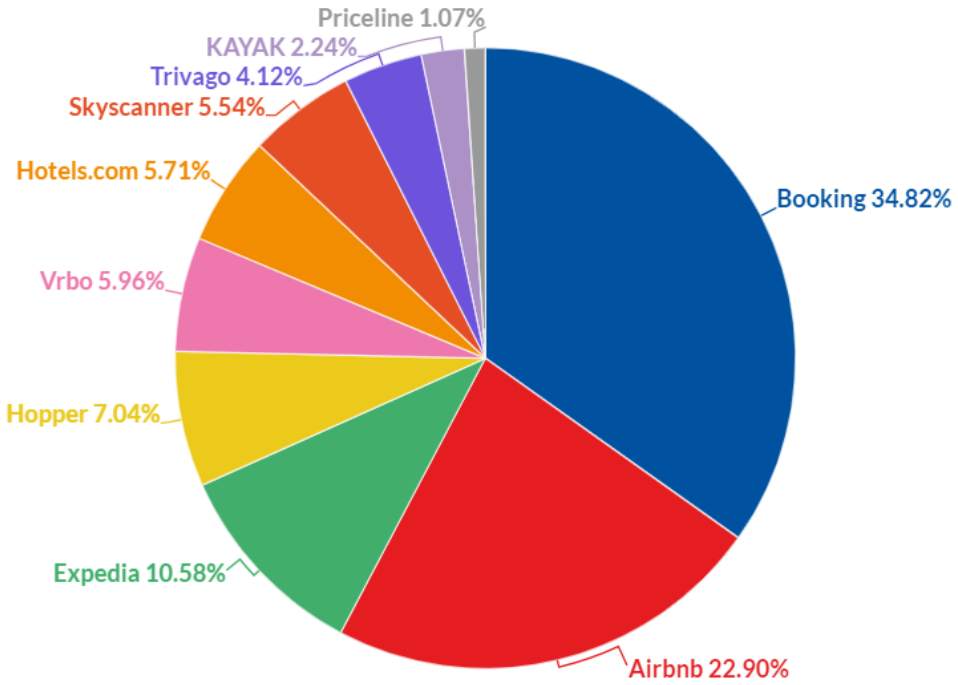

Рис. 1.2. Частка Airbnb на світовому ринку

Джерело: [19]

Основна причина популярності компанії полягає в тому, що компанія має здатність адаптуватись до конкретних потреб споживачів, спираючись на те, що право власності вже не є пріоритетним у споживчій поведінці. Послуги Airbnb забезпечують ефективність та зручність взаємодії між власниками житла та орендарями завдяки системі технологічних рішень, що дозволяє швидко та безпечно здійснювати бронювання та оплату житла, отримувати зворотній зв’язок, бачити оцінки інших користувачів та оцінювати якість послуги.

Варто зазначити, що компанія успішно використовує сучасні технології та активно розвиває бренд у соціальних мережах, забезпечуючи просування Airbnb на нові ринки та залучаючи нових користувачів. Це дозволило компанії вирости з невеликого стартапу до міжнародного бренду, який став невід’ємною частиною індустрії туризму, готельного сектору та послуг оренди.

Ще одним вагомим фактором успіху компанії Airbnb є наявність персоналізованого підходу до кожного клієнта. Прагнення компанії створити унікальний досвід проживання дозволяє відвідувачам вибрати місце проживання відповідно до їхніх конкретних потреб і бажань, з додатковим бонусом доступу до місцевих господарів для отримання інформації про околиці. Крім того, Airbnb надає низку послуг, таких як обслуговування та екскурсії, які дозволяють гостям повністю зануритися в місцеву культуру. Це поєднання індивідуального житла та незабутніх вражень від подорожей зробило Airbnb популярним вибором для мандрівників.

Усі ці чинники роблять Airbnb одним з найбільш перспективних та інноваційних лідерів в індустрії гостинності завдяки ефективному поєднанні передових технологій з практичними потребами споживачів, пропонуючи комфортні та зручні варіанти житла для мандрівників по всьому світу.

Вивчення використання платформ спільної економіки, таких як Airbnb, у різних географічних місцях має велике значення з різних причин. Це дозволяє проводити поглиблені наукові дослідження, щоб краще зрозуміти зв'язок між географією міст і цією економічною моделлю, особливо в епоху, коли технології та легкість подорожей дозволяють відвідувати різні країни.

Обираючи компанію Airbnb для дослідження аналітики клієнтів за допомогою сегментації, можна виділити декілька умов, за яких ця компанія є цікавою та релевантною для подальшого дослідження:

- інноваційний підхід до надання послуг оренди, який забезпечує клієнтам унікальні враження. Вивчаючи клієнтську базу Airbnb, можна отримати цінну інформацію про те, як ці інноваційні стратегії впливають на споживачів і як компанія адаптується до потреб різних сегментів клієнтів;
- вихід компанії на глобальний ринок, що дозволяє аналізувати різноманітність та особливості клієнтів з різних культур, регіонів та географічних локацій;

- доступність статистичних даних, створених на основі інформації про клієнтську поведінку;
- значний вплив на галузь: компанія не тільки змінила підхід до туризму та короткострокової оренди, але також має великий вплив на готельну індустрію та традиційні туристичні послуги.

Зокрема, варто також зазначити, що Airbnb в рамках проєкту United for Ukraine створив програму з надання безкоштовного тимчасового житла українським біженцям в Європі, плануючи надати прихисток понад 100 тисячам осіб [20].

Враховуючи всі ці фактори, вибір компанії Airbnb для дослідження аналітики клієнтів на основі сегментації є доцільним та обґрунтованим. Вона надає широкі можливості для отримання цінного досвіду, який може бути використаний для розвитку стратегій та вдосконалення послуг у сфері гостьового сектору.

1.4. Сегментація споживачів

Історія успіху Airbnb є свідченням потужності прийняття рішень на основі даних. Airbnb завжди була прозорою щодо використання даних для створення нових пропозицій продуктів, покращення послуг і збільшення прибутку за допомогою інноваційних маркетингових ініціатив. За словами Райлі Ньюмана, колишнього керівника відділу обробки даних Airbnb, компанія розглядає дані як голос клієнта, а науку про дані як інтерпретацію цього голосу. Використовуючи дані, щоб отримати розуміння поведінки та вподобань клієнтів, Airbnb зміг адаптувати свої пропозиції відповідно до потреб своїх користувачів, що призвело до неймовірного зростання та успіху [21].

Аналіз клієнтської бази компанії Airbnb має вирішальне значення для підвищення ефективності маркетингових кампаній і підвищення стандартів обслуговування клієнтів. Результати аналізу за допомогою сегментації клієнтів можуть використовуватись для виокремлення визначальних рис конкретних

сегментів клієнтів та застосовуються маркетологами і брендами для визначення найефективніших кампаній, пропозицій або продуктів для подальшого впровадження в маркетингову стратегію відповідно до вподобань певних сегментів споживачів. Крім того, компанії можуть використовувати аналіз сегментації клієнтів, щоб визначити потенційну цінність певних сегментів шляхом ретельного вивчення прогнозованої майбутньої вартості, середньої вартості замовлення, розподілу лояльності та інших факторів.

Варто зазначити, що базисом будь-якої стратегії, орієнтованої на клієнтів, є сегментація клієнтів. Це фундаментальний крок до створення моделей для аналітики поведінки клієнтів, розробки стратегій їх утримання та створення персоналізованого досвіду для кожного клієнта.

Сегментацією клієнтів (сегментацією ринку) називають процес ідентифікації груп клієнтів зі спільними характеристиками або категоріями. Ця класифікація спрямована на оптимізацію маркетингу для кожної групи, надаючи окремим клієнтам найбільш доречну та персоналізовану інформацію, максимізуючи їх цінність для компанії. Оскільки компанія розширюється, разом з нею розширюється і її клієнтський портфель. Однак із збільшенням кількості клієнтів ймовірність того, що вони матимуть ознаки якогось одного профілю клієнта, зменшується, що ускладнює управління та задоволення унікальних потреб клієнтів. Одним з способів вирішення цієї проблеми є застосування сегментації клієнтів. Поділяючи клієнтів на певні групи, підприємства можуть краще пристосовувати свої клієнтські стратегії для задоволення індивідуальних потреб кожної групи [20].

На сьогодні питання сегментації користувачів послуг компанії Airbnb неодноразово досліджували різні науковці, використовуючи різні методи сегментації та аналізуючи виокремлені сегменти. В таблиці нижче розглянуто напрями досліджень основних робіт, які мають значний вплив на розуміння сегментації користувачів Airbnb:

Таблиця 1.1

Відомі дослідження сегментації користувачів Airbnb

Назва роботи	Опис дослідження	Використані методи
Giovanni Quattrone, Natalia Kusek, Licia Capra. A global-scale analysis of the sharing economy model – an AirBnB case study.	Головна мета дослідження цієї роботи стосується лінгвістичного аналізу, аналізу настроїв, вивчення подібності та відмінності впровадження Airbnb у всьому світі.	Обробка природної мови, виконана на основі алгоритму Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM) та латентного розподілу Діріхле (LDA).
Christoph Lutz, Gemma Newlands. Consumer segmentation within the sharing economy: The case of Airbnb.	Дослідження сегментацію споживачів в рамках єдиної платформи економіки спільного використання: Airbnb. Використовуться підхід змішаних методів, як з кількісним опитуванням, так і з якісним аналізом вмісту оголошень Airbnb, порівнюється різні типи оренди житла.	Лінійна регресія для пояснення впливу незалежних змінних на частоту використання різних типів житла, кількісне опитування користувачів Airbnb, а також проведення якісного аналізу вмісту оголошень Airbnb.

Продовження табл. 1.1

Sowmya Vivek. Clustering algorithms for customer segmentation.	Продемонстровано концепцію сегментації набору даних клієнта з сайту електронної комерції за допомогою кластеризації, практична реалізація виконана мовою програмування python.	Алгоритм кластеризації K-means з використанням WCSS.
Del Chiappa, G., Sini, L., Atzeni, M. A motivation-based segmentation of Italian Airbnb users: an exploratory mixed method approach.	У статті застосовано дослідницький підхід до змішаного методу в Італії. Зокрема, проведено якісне дослідження, засноване на опитуваннях, якісні результати були використані для інформування, доповнення існуючої літератури, розробки інструменту опитування для збору даних.	Факторно-кластерний аналіз шляхом проведення опитування та якісного і кількісного дослідження його результатів.

Продовження табл. 1.1

Anjana S. Customer Segmentation using Cluster Analysis. Medium.	Розглянуто модель сегментації клієнтів. Зокрема, використано алгоритми кластеризації K-means та ієрархічну кластеризацію, проаналізовано та візуалізовано вхідні дані та результати.	Роздільна ієрархічна кластеризація, K-means.

Джерело: створено автором на основі [21],[22],[23],[24],[25]

Отже, розглянуті дослідження базуються на як на кількісному, так і на якісному аналізі. Основними моделями кластеризації є K-means та ієрархічна кластеризація.

Зокрема, існує багато характеристик і факторів, за якими можна проводити сегментацію. Розглянемо найпоширеніші з них:

- демографічна – містить дані про загальні характеристики, такі як вік, стать, дохід, місце проживання, освіта, рід занять, сімейний стан, релігія, національність тощо. Розширені моделі сегментації зазвичай базуються на визначених демографічних групах;
- географічна – поділ клієнтів на основі їхнього географічного місцезнаходження, рельєфу чи клімату;
- психографічна – розуміння особистостей, поглядів, цінностей та інтересів користувачів. Психографічні дані можуть охоплювати низку тем, включаючи риси особистості, збереження навколишнього середовища, а також такі хобі;
- поведінкова – групування відповідно до поведінки споживачів;

- сегментація на основі вартості – визначення цінності клієнта за сукупною сумою, яку вони витрачають на придбання послуг, тобто на основі цінності, яку вони приносять компанії;
- технологічна сегментація – застосовується при процесі поділу користувачів на основі технології, яку вони використовують, або їх стеку технологій [26].

Методологія групування клієнтів і прогнозування їхніх купівельних звичок є надзвичайно різноманітною, включаючи кластеризацію, дерева рішень, узагальнені моделі правил асоціації та нейронні мережі. Хоча пошук цих сегментів, безумовно, є викликом для ринку, найскладнішим завданням є розробка ефективної бізнес-стратегії для цих груп. Найцінніша інформація, яку можна виокремити при даному дослідженні, це виявити, чи існують якісь значні відмінності між цими групами та як можна скористатись цими відмінностями. Зрештою, метою сегментації ринку є надання компаніям необхідної інформації для встановлення оптимальної взаємодії з потенційними споживачами та визначення тих сегментів, на які можна легко орієнтуватися. Завдяки покращенню комунікації та ефективнішій клієнтоорієнтованості можна збільшити прибутки та підвищити забезпечення потреб користувачів. Базуючись на ресурсах, можна вирішити, чи спиратись на багато сегментів, чи лише на обмежену кількість в управлінні компанією.

Висновки до розділу 1

Отже, перший розділ описує взаємозв’язок між платформами спільного споживання і цифровізацією економіки на прикладі галузі оренди нерухомості. Зокрема, було проаналізовано основні концепції та принципи моделі спільного споживання. Виявлено, що економіка спільного споживання привела до змін у споживацькій поведінці та вподобаннях, в тому числі і в сфері туризму.

Також детально розглянуто компанію Airbnb як представника цифрової платформи економіки спільного споживання. Було виявлено, що Airbnb відіграє

значну роль у сфері короткострокової оренди житла та є популярною серед охочих винаймати житло. Останній підрозділ присвячено аналіз наявних досліджень та уявленню про різні типи клієнтів Airbnb та їхні поведінкові характеристики.

Отже, аналітика клієнтів компанії Airbnb за допомогою сегментації на основі наявних даних може бути ефективним інструментом для розуміння потреб, уподобань та поведінки різних клієнтських сегментів. Це дозволяє компанії вдосконалювати свої стратегії та пропонувати більш персоналізовані послуги.

РОЗДІЛ 2. МЕТОДОЛОГІЧНІ ЗАСАДИ СЕГМЕНТАЦІЇ

2.1. Особливості та підходи до сегментації

Сегментація клієнтів - одна з ключових областей аналітики клієнтів, яка допомагає ідентифікувати різні групи користувачів відповідно до їх поведінки. Визначення основних сегментів допомагає зрозуміти, націлити і краще спілкуватися з існуючими клієнтами і залучати до бізнесу нових, що є основою успішного маркетингу [27].

Перехід до цифрового бізнесу та оцифрування економіки є важливим етапом у розвитку електронної комерції. Поєднання цифрових технологій у бізнесі спонукало до збільшення кількості даних про клієнтів, які генеруються з різних джерел, включаючи веб-сайти, соціальні мережі, численні програмні застосунки та інші цифрові платформи. Ця компіляція даних створює величезну кількість знань про взаємодію та поведінку клієнтів. Аналіз цих даних, відомий як клієнтська аналітика, передбачає їх перевірку та вивчення для кращого розуміння переваг, потреб та поведінки клієнтів. Згодом ці знання можна використати для створення нових продуктів і послуг чи оновлення існуючих, відповідно до нових вимог.

Інтеграція цифрових та інформаційних технологій у різні бізнес-операції дозволила аналізувати поведінку клієнтів у режимі реального часу. Це дозволяє підприємствам швидко визначати закономірності та тенденції, що сприяє підвищенню ефективності процесів.

Використовуючи зібрані дані, підприємства можуть приймати рішення на основі фактичної інформації щодо маркетингових стратегій, розробки продукту та обслуговування клієнтів. Ретельно вивчаючи дані клієнтів, компанії отримують глибше розуміння бажань та уподобань своїх споживачів, у свою чергу, дозволяючи їм персоналізувати свої продукти та послуги, щоб краще задовольнити ці переваги. Аналіз історії покупок дає цінну інформацію про те, які продукти користуються попитом серед певних сегментів споживачів, що дозволяє компанії відповідно коригувати свої пропозиції продуктів або послуг.

Крім того, відстеження взаємодії в соціальних мережах дозволяє компаніям отримати цінну інформацію про вподобання клієнтів, які потім можна використовувати для покращення маркетингових повідомлень і покращення стратегій обслуговування клієнтів.

Всі перелічені вище операції та сценарії можна реалізувати за допомогою проведення сегментації клієнтів, зокрема, за допомогою сегментації на основі кластеризації бази даних споживачів, що є практичною основою даної дипломної роботи.

Сегментація споживацького ринку — це розуміння рушійних факторів вибору, а також варіантів, які доступні споживачам. Відомим і широко використовуваним методом застосування сегментації ринку є кластерний аналіз, який відноситься до класу методів неконтрольованого навчання та використовується для класифікації осіб на групи [28].

Варто також згадати, що в різних джерелах часто плутають поняття сегментація та кластеризація. Хоча вони разом відносяться до області аналітики клієнтської бази, дані терміни не є взаємозамінними, оскільки кластеризація має на меті розподіл точок даних, які згруповані на основі певних статистичних особливостей. Для визначення подібності в даних споживачів кластерний аналіз використовує алгоритми машинного навчання (ML), які перевіряють клієнтські дані, відзначають подібності, і об’єднують точки даних, які відповідають за клієнтську статистичну інформацію у кластери на основі моделей їхньої поведінки. Сегментація, в свою чергу, є більш широким поняттям, оскільки охоплює поділ кластерів на сегменти з добре описаними характеристиками, що включають поведінку користувачів, ціноутворення, політику тощо [29].

Важливим аспектом є те, що у контексті сегментації клієнтів аналіз кластеризації клієнтів використовує математичну модель для ідентифікації кластерів клієнтів, що мають спільні риси, шляхом визначення найдрібніших розбіжностей між клієнтами в кожному кластері. Ці однорідні групи зазвичай називають «архетипами клієнтів» або «персонами». Результатом кластерного

аналізу є точне сегментування споживачів, яке застосовується для досягнення більш ефективного маркетингу за допомогою персоналізації стратегій [30].

Процес неконтрольованого машинного навчання передбачає надання комп'ютеру вказівок використовувати дані, які не мають міток і класифікації, а потім дозволяє алгоритму працювати з цими даними без будь-якого контролю. З огляду на те, що машина не була попередньо навчена цьому конкретному набору даних, її метою є систематичне впорядкування несортованих даних шляхом виявлення подібностей, шаблонів та відхилень.

Отже, кластеризація — це техніка, призначена для поділу генеральної сукупності або набору точок даних на кілька груп. Це робиться для того, щоб точки даних у кожній групі мали вищий ступінь схожості одна з одною, водночас відрізняючись від точок даних в інших групах. За своєю суттю кластеризація — це метод сегментації елементів на основі ступеня їхньої схожості та відмінності [31].

Незважаючи на переваги та обмеження всіх методів, K-Means, тип алгоритму розділеної кластеризації, зазвичай показує себе найкраще в реалізації, тому використовується як основний метод у подальшій роботі. Проте варто також доповнити дослідженні іншими методами, які можуть в тій чи іншій мірі показувати реультати кластеризації клієнтів.

Саме тому в даній дипломній роботі сегментація клієнтів Airbnb буде проведена за допомогою наступних методів кластеризації: K-Means, DBSCAN, BIRCH. Такі методи кластеризації були обрані з метою отримання різних підходів до сегментації клієнтів Airbnb. Кожен з методів кластеризації має власний унікальний набір атрибутів, що дозволяє отримати різні результати.

Порівнюючи методи кластеризації, ми можемо оцінити їхню ефективність у проведенні сегментації клієнтів Airbnb, порівнюючи результати кожного методу та визначаючи будь-які можливі недоліки чи переваги, які може мати кожен метод.

В цілому, застосування цих методів кластеризації дозволить більш змістовно зрозуміти поведінку та потреби клієнтів Airbnb, створенню більш ефективних стратегій та покращенню задоволеності користувачів.

Зокрема, також можна виокремити наступний план впровадження методів кластеризації:

- імпортуємо необхідні дані та бібліотеки;
- створюємо основні візуалізації даних і робимо попередній дослідницький аналіз даних;
- очищуємо, об'єднуємо та аналізуємо дані при підготовці до використання методів кластеризації;
- використовуючи різні підходи, визначаємо відповідну кількість кластерів;
- навчаємо та реалізовуємо алгоритми моделі ML;
- знаходимо і ідентифікуємо кластери, що містять спільні характеристики;
- реалізуємо візуалізацію кластерного аналізу;
- описуємо результати та порівнюємо різні методи кластеризації.

Отже, аналіз кластерів може розкрити певні групи клієнтів, які мають спільні потреби та допомогти виявити можливі напрями у вдосконаленні послуг, спрямованих на задоволення цих потреб. Крім того, ідентифікація сегментів клієнтів може також вказати на потенційні нові ринкові можливості або сегменти, покращення показників яких варто розглянути.

2.2. Моделі кластеризації

Розглянемо теоретичний аспект реалізації наступних моделей кластеризації: K-Means, DBSCAN, BIRCH.

2.2.1. K-Means

K-Means - один з найпопулярніших алгоритмів кластеризації, який застосовується при вирішення широкого спектру завдань. Кластеризація з використанням K-Means є алгоритм неконтрольованого машинного навчання,

метою якого є групування набору даних в k попередньо визначених кластерів з точками, які будуть близькі в межах одного скупчення і віддалені з усіх інших скупчень. Відстань зазвичай вимірюється евклідовою відстанню [32].

Алгоритм K-Means кластеризує дані, намагаючись розділити вибірки на n груп однакової дисперсії, мінімізуючи критерій, відомий як інерція або сума квадратів у межах кластера. Для реалізації цього алгоритму обов’язково потрібно вказати кількість кластерів, яку можна визначити за допомогою «методу ліктя».

Модель отримує як вхідні дані кількість кластерів для формування, а також набір векторів, що представляють спостереження. Також створює набір центроїдів, ідентичних для кожного з k кластерів. Вектору спостереження присвоюється номер кластера, який пов'язаний з центроїдом кластера або індексом центроїда, який є найближчим до нього [33].

Алгоритм k-means ділить набір із N вибірок на K непересічних кластерів, кожен з яких описується середнім значенням вибірок у кластері. Тобто алгоритм має на меті вибрати центроїди, які мінімізують інерцію або критерій суми квадратів у кластері, та обраховується за наступною формулою:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (\| x_i - \mu_j \|^2)$$

, (2.1)

де n – кількість записів;

C – кількість кластерів;

x_i - i-та точка даних;

μ_j - j-тий центроїд кластера [34].

Основними перевагами застосування моделі k-means є відносно проста реалізація та швидкість виконання. Такий алгоритм найкраще підійде для сегментування великої кількості даних, оскільки є менш обчислювально інтенсивним.

Попри значні переваги, алгоритм також має недоліки, серед яких: самостійне визначення кількості кластерів, чутливість алгоритму до викидів, висока ймовірність неточності результатів [35].

2.2.2. DBSCAN

DBSCAN (density-based spatial clustering of applications with noise) – відома модель кластеризації даних, яка часто використовується в добуванні даних та машинному навчанні.

DBSCAN збирає точки, близькі один до одного на основі сукупності точок (наприклад, в двовимірному просторі) на основі вимірювання відстані (часто евклідової відстані) і мінімальної кількості точок. Крім того, він визначає ділянки в районах з низькою щільністю як викиди [36].

Даний алгоритм базується на 2 параметрах:

1. Eps: визначає околиці навколо точки даних, тобто якщо відстань між двома точками менша або дорівнює eps, ці точки даних вважаються сусідніми. Якщо значення eps вибрано занадто мале, велика частина даних буде вважатися викидами даних. Якщо значення параметра занадто велике, кластери об’єднаються, і більшість точок потрапить в однакові кластери. Один із прийнятних способів знайти значення eps – це застосувати графік k-відстаней.
2. MinPts: мінімальна кількість сусідів (точок даних) у радіусі eps. Тобто обсяг набору даних прямо корелює із значенням MinPts. Зазвичай, мінімальний MinPts можна отримати з кількості вимірів D у наборі даних таким чином: MinPts >= D+1. Обов’язкова умова: мінімальне значення MinPts має бути не менше 3.

Зокрема, на малюнку нижче розташована ілюстрація поділу точок бази даних за допомогою DBSCAN.

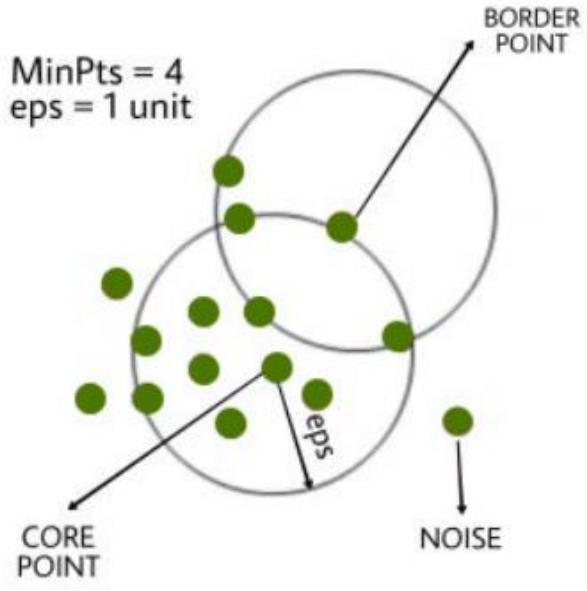

Рис. 2.1. Схема поділу бази даних при моделюванні DBSCAN
Джерело: [37]

Як можна виділити, у цьому алгоритмі є 3 типи точок даних: core point (точка вважається основною, якщо вона має більше ніж MinPts балів у межах eps), border point (точка, яка має менше ніж MinPts в межах eps, але вона знаходиться поблизу основної точки) та noise (викиди або точки, які не є основними і не межують з будь-якими іншими точками) [37].

2.2.3. BIRCH

Balanced iterative reducing and clustering using hierarchies (BIRCH) — це алгоритм кластеризації, який може кластеризувати великі набори даних, спочатку генеруючи невеликий підсумок великого набору даних, який зберігає якомога більше інформації. Цей менший підсумок потім кластеризується замість кластеризації більшого набору даних. BIRCH часто використовується для доповнення інших алгоритмів кластеризації шляхом створення підсумку набору даних, який тепер може використовувати інший алгоритм кластеризації. Однак BIRCH має один серйозний недолік – він може обробляти лише метричні атрибути [38].

Робота моделі Birch полягає в побудові дерева Clustering Feature Tree (CFT) для вхідних даних. Дані, в свою чергу, стискаються з втратами даних до набору

вузлів Clustering Feature (CF Nodes). CF Nodes мають ряд підкластерів, які мають назву Clustering Feature subclusters (CF Subclusters), CF Subclusters розташовані в нетермінальних CF Nodes, та можуть мати дані вузли як дочірні.

Підкластери CF містять необхідну інформацію для кластеризації, що запобігає необхідності зберігати всі вхідні дані в пам’яті. Ця інформація включає: кількість зразків у субкластері, лінійну суму (n-вимірний вектор, що містить суму всіх вибірок), суму в квадраті, центроїди, квадрат норми центроїдів.

Також варто зазначити, що алгоритм BIRCH має два параметри: поріг і коефіцієнт розгалуження. Коефіцієнт розгалуження обмежує кількість субкластерів у вузлі, а поріг обмежує відстань між вхідною вибіркою та існуючими субкластерами [39].

2.3. Огляд бази даних

При виконанні аналізу та подальшій сегментації користувачів важливою запорукою отримання релевантних результатів є використання інформації з надійних джерел, що мають оновлену статистику. Неточні, застарілі або недостовірні дані можуть призвести до помилкових результатів. Таким чином, застосування надійних і актуальних статистичних даних є важливим аспектом моєї дипломної роботи, що забезпечує точність і надійність висновків.

Дослідження, проведене в даній дипломній роботі, виконане на основі даних, які представлені на вебсайті «Inside Airbnb» [40]. Даний ресурс починав свою діяльність як активістський проект для моніторингу ролі туристичної оренди. Протягом багатьох років це не лише допомогло місцевим господарям, які прагнули стати частиною спільноти Airbnb, а й дослідникам і науковцям. Сайт публікує статистичну інформацію про оголошення на ресурсах компанії Airbnb, та містить наступні дані: кількість доступних оголошень, безпечні райони для розміщення, орієнтовний місячний дохід, середня кількість забронюваних ночей, ціна оренди за ніч, рівень заповнюваності, кількість оглядів тощо [41].

Статистика, наведена на Inside Airbnb, охоплює інформацію для кожного оголошення Airbnb з даними для певного регіону чи міста. Зокрема, для подальших розрахунків було обрано наступні міста: Амстердам, Афіни, Барселона, Берлін, Лісабон, Лондон, Мадрид, Париж, Рим, Відень. Такий вибір міст спрямований на дослідження загальної тенденції взаємодії користувачів Airbnb у межах популярних європейських міст та проведення сегментації користувачів саме на європейському ринку.

Завантажені дані містять інформацію, яка сформована в період з 9.03.2023 по 27.03.2023 та охоплює наявні в доступності записи за весь час реєстрації оголошень на ресурсах Airbnb.

Для реалізації практичної частини даної дипломної роботи я використовую IDE Google Colaboratory (Google Colab). Це хмарний сервіз розроблений Google Research, який дозволяє писати вихідний код в редакторі і запускати його безпосередньо з браузера. Також реалізація програмної частини виконана на такій мові програмування, як Python. Такий вибір обґрунтований тим, що Python містить значну кількість бібліотек, орієнтованих на виконання машинного навчання, аналізу даних, візуалізації тощо [42].

Перш ніж почати ознайомлення з базою даних, необхідно завантажити бібліотеки, за допомогою яких буде проводитись весь наступний аналіз. Зокрема, для аналізу даних застосовуються такі бібліотеки, як Pandas і Numpy, а для візуалізації даних – бібліотеки Matplotlib і Seaborn.

Варто зазначити, що вхідні дані мають вигляд 10 таблиць у файловому форматі csv, де в кожній з таблиць міститься інформація про одну певну країну. Після завантаження цього набору даних доцільно переконатись, що бази даних мають однакові атрибути, що можна побачити на малюнку нижче:

```
(6998, 18)
(11382, 18)
(15655, 18)
(12049, 18)
(20097, 18)
(75241, 18)
(21239, 18)
(56726, 18)
(24924, 18)
(12525, 18)
```

Рис. 2.2. Параметри бази даних

Джерело: розрахунки автора

Всі набори мають однакову кількість змінних, при знаходженні всіх унікальних індексів маємо наступні значення:

```
Index(['id', 'name', 'host_id', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'latitude', 'longitude', 'room_type', 'price',
       'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'calculated_host_listings_count',
       'availability_365', 'number_of_reviews_ltm', 'license'],
      dtype='object')
```

Рис. 2.3. Змінні бази даних

Джерело: розрахунки автора

Отже, набір даних містить наступні змінні:

Таблиця 2.1

Опис змінних

№	Змінна	Опис	Тип
1	id	унікальний ідентифікатор для оголошення	integer
2	name	назва оголошення	string

Продовження табл. 2.1

3	host_id	унікальний ідентифікатор для хоста/користувача	integer
4	host_name	ім'я хоста	string
5	neighbourhood_group	група околиць (геокодована з використанням широти та довготиьоколиць, як визначено відкритими або загальнодоступними цифровими шейп-файлами)	String
6	neighbourhood	назва околиць	string
7	latitude	широта розміщення житла (на основі проекції Всесвітньої геодезичної системи (WGS84))	float
8	longitude	довгота розміщення житла (на основі проекції Всесвітньої геодезичної системи (WGS84))	float
9	room_type	різновиди житла (цілий будинок, окрема кімната, спільна кімната, ціле місце)	string
10	price	щоденна ціна в місцевій валюті.	integer
11	minimum_nights	мінімальна кількість ночей для розміщення	integer
12	number_of_reviews	кількість відгуків, які має одне оголошення	integer
13	last_review	дата останнього перегляду оголошення	date

Продовження табл. 1.1

14	reviews_per_month	кількість відгуків за весь період існування оголошення	Float
15	calculated_host_listings_count	кількість списків, які має хост	Integer
16	availability_365	доступність оголошення протягом року	integer
17	number_of_reviews_ltm	кількість переглядів оголошення (за останні 12 місяців)	integer
18	license	ліцензія/дозвіл/реєстраційний номер	string
19	area	назва країни	string

Джерело: створено автором на основі [44]

При перевірці відповідності типів даних можна побачити, що дані відповідають необхідним вимогам. У базі даних є 19 змінних, серед яких 8 цілочисельних, 7 текстових, 3 дійсних та 1 змінна, що має формат дати. Деякі змінні неінформативні або ж непотрібні для виконання практичної частини, тому в подальшому не будуть включені в остаточно сформовану загальну базу даних.

Висновки до розділу 2

Даний розділ присвячений визначенню концепції сегментації користувачів Airbnb та ознайомленню з теоретичним аспектом кластеризації. Зокрема, одним із важливих компонентів аналітики клієнтів є сегментація, яка передбачає виявлення груп клієнтів з чіткими моделями поведінки. Завдяки цьому компанія Airbnb може отримати уявлення про потреби та бажання своїх клієнтів і ефективно взаємодіяти з користувачами послуг.

Також визначено теоретичні особливості методів, впровадження яких буде реалізовано в 3 розділі. Для цього було обрано наступні такі методи: K-Means, DBSCAN, OPTICS, BIRCH, ward hierarchical clustering.

Зокрема, в межах даного розділу розглянуто базу даних, яка буде застосовуватись в дослідженні та описано основні її критерії. Програмна частина дипломної роботи також буде виконана на мові програмування Python з використанням бібліотек для реалізації алгоритмів кластеризації.

В наступному розділі буде розглянуто аспекти практичного застосування описаних алгоритмів кластеризації. Зокрема, перший підрозділ третього розділу буде присвячено аналізу даних, їх очищенню та підготовці до безпосереднього застосування. Згодом розглянемо реалізацію моделей кластеризації та порівняємо їх за результативністю.

РОЗДІЛ 3. ПРАКТИЧНЕ ЗАСТОСУВАННЯ МЕТОДІВ СЕГМЕНТАЦІЇ

3.1. Аналіз даних

Важливою передумовою проведення сегментації задля отримання більш точних та покращених результатів є аналіз даних та їх очищення.

Перевіримо кількість унікальних значень для змінної «room_type» та побудуємо графік, що показує відношення різних типів орендованих приміщень у різних містах.

Наступним кроком буде очищення бази даних та створення нових таблиць шляхом відкидання нерелевантних змінних. Нові створені бази даних для кожної з країн містять наступні 10 змінних: id, neighbourhood, latitude, longitude, room_type, price, minimum_nights, number_of_reviews, reviews_per_month, availability_365.

Додаємо нову змінну «Area», що позначає назву країни в певному наборі даних.

Далі, з допомогою функції concat() створюємо нову базу даних "df_concat", яка містить в собі статистичні дані 10 країн, має 11 змінних та охоплює 256836 записів. Найбільше оголошень доступно для Лондону, Парижу та Риму, що пояснюється високою популярністю цих міст серед туристів.

```
London       75241
Paris        56726
Rome         24924
Madrid       21239
Lisbon       20097
Barcelona    15655
Vienna       12525
Berlin       12049
Athens       11382
Amsterdam     6998
Name: Area, dtype: int64
```

Рис. 3.1. Кількість записів

Джерело: розрахунки автора

При перевірці наявності відсутніх значень було виявлено, що змінна «reviews_per_month» має 19,66% пропущених значень, які усунуті шляхом заміни відсутніх значень середнім значенням для цього стовпця.

	percent_missing
id	0.000000
neighbourhood	0.000000
latitude	0.000000
longitude	0.000000
room_type	0.000000
price	0.000000
minimum_nights	0.000000
number_of_reviews	0.000000
reviews_per_month	19.662353
availability_365	0.000000
Area	0.000000

Рис. 3.2. Перевірка пропущених значень

Джерело: розрахунки автора

Наступним кроком буде знаходження рядків, де ціна оренди дорівнює нулю. Враховуючи, що таких рядків не так багато (62), я заміню ці числа середнім значенням, враховуючи, що стовпець ціни містить кілька екстремальних викидів. Отримуємо, що середня ціна оренди становить 98 євро.

	id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	availability_365
count	2.568360e+05	256836.000000	256836.000000	256836.000000	256836.000000	256836.000000	256836.000000	256836.000000
mean	2.059630e+17	46.770303	3.409107	162.481257	24.427790	33.513043	1.181030	135.612566
std	3.264393e+17	5.089295	7.736780	819.182455	79.269166	68.489871	1.379283	131.680204
min	2.737000e+03	37.950550	-9.487890	1.000000	1.000000	0.000000	0.010000	0.000000
25%	1.902862e+07	41.400564	-0.177750	60.000000	1.000000	1.000000	0.280000	0.000000
50%	3.800288e+07	48.859830	2.177690	98.000000	2.000000	8.000000	1.000000	95.000000
75%	6.092720e+17	51.501240	4.889622	160.000000	5.000000	33.000000	1.290000	261.000000
max	8.556783e+17	52.656110	23.780220	100180.000000	1125.000000	2524.000000	111.590000	365.000000

Рис. 3.3. Статистика змінних

Джерело: розрахунки автора

При перевірці повторюваних записів не було виявлено, тобто всі рядки бази даних є унікальними. Проте, для уникнення помилок в подальших обчисленнях та уніфікації даних перетворюємо формат змінних «Price», «Minimum Nights»,

«Number of Reviews», «Availability (365)» на тип даних float. Також за допомогою методу get_dummies() змінюємо категоріальні змінні в числовий формат.

Також побудуємо графік кореляції змінних, який має наступний вигляд:

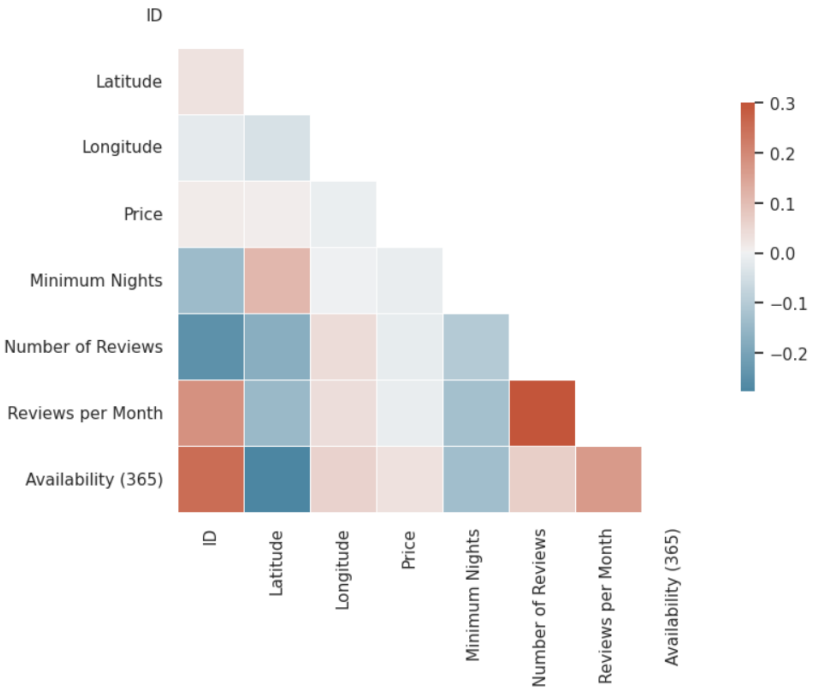

Рис. 3.4. Графік кореляції

Джерело: розрахунки автора

Тимчасово видаливши змінні, які не будуть брати участь у реалізації методів сегментації, отримуємо остаточну базу даних, яка має наступний вигляд:

	Price	Minimum Nights	Number of Reviews	Reviews per Month	Availability (365)
0	69.0	3.0	322.0	1.90	44.0
1	106.0	1.0	339.0	2.14	0.0
2	143.0	3.0	248.0	1.82	14.0
3	76.0	2.0	476.0	3.12	79.0
4	56.0	2.0	618.0	4.23	69.0

Рис. 3.5. Остаточна база даних

Джерело: розрахунки автора

Для безпосереднього використання бази даних масштабуємо змінні за допомогою функції StandardScaler() та створюємо масив даних, який можна застосовувати при моделюванні.

```
array([[-0.11411553, -0.27031737,  4.21211944,  0.52126435, -0.69572147],
       [-0.06894846, -0.29554791,  4.46033181,  0.69526812, -1.02986498],
       [-0.02378139, -0.27031737,  3.1316656 ,  0.46326309, -0.92354659],
       ...,
       [-0.11533626, -0.29554791, -0.48931481,  0.        ,  1.68884814],
       [-0.07627285, -0.29554791, -0.48931481,  0.        ,  1.65087729],
       [ 0.10805763, -0.29554791, -0.48931481,  0.        , -0.96151744]])
```

Рис. 3.6. Перетворення даних на масив

Джерело: розрахунки автора

3.2. Реалізація моделей кластеризації

Після очищення і формування остаточної бази даних переходимо до безпосередньої реалізації методів сегментації, теоретичний аспект яких описаних в 2 розділі дипломної роботи.

3.2.1. K-Means

Розпочнемо огляд результатів практичної частини дослідження з реалізації класифікаційної моделі K-Means.

Першим кроком для побудови моделі є знаходження оптимальної кількісті кластерів за допомогою методу «ліктя». Для цього необхідно створити графік, який відображає кількість кластерів на осі X і WCSS для кожного номера кластера на осі Y. Зі збільшенням кількості кластерів графік має тенденцію до спаду, а так званою «точкою ліктя» є точка, де сповільнюється спадання графіка. Цей показник в поєднанні з конкретними знаннями бізнес-вимоги може бути використаний для прийняття рішення про оптимальну кількість кластерів.

Відповідно до зображення нижче, можна зробити висновок, що кількість кластерів дорівнює 5, оскільки саме у цій точці відбувається різкий спад кількості кластерів.

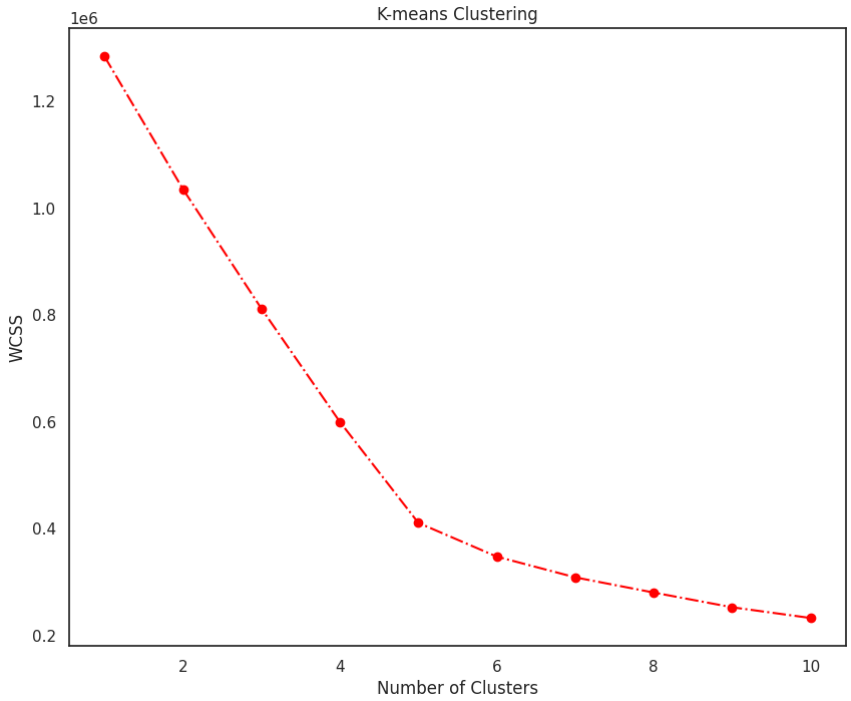

Рис. 3.7. Графік пошуку кількості кластерів методом «ліктя».

Джерело: розрахунки автора

Отже, грунтуючись на отриманій кількості кластерів наступним кроком буде побудова власне моделі кластеризації шляхом навчання моделі та створення об’єкту «kmeans» із 5 кластерами на основі обробленої бази даних.

Результатом проведення кластерного поділу отримано 5 сегментів, які містять інформацію про 5 кластерів та 5 змінних, які були основою для навчання моделі кластеризації.

	Price	Minimum Nights	Number of Reviews	Reviews per Month	Availability (365)
0	189.598645	7.918400	19.297091	1.091433	278.531945
1	141.417868	8.023921	16.943754	0.779943	31.703132
2	111.562282	368.526621	7.054404	0.515763	54.389949
3	120.033890	2.947052	185.714698	3.906881	163.195294
4	73976.541667	50.000000	41.125000	0.794712	104.625000

Рис. 3.8. Результат сегментації моделі K-Means

Джерело: розрахунки автора

В результаті обрахунків було отримано 5 кластерів, сформовані на основі 5 характеристик: ціна, мінімальна кількість ночей, кількість відгуків, відгуки

протягом місяця, доступність оренди протягом року. Інтерпретувати отримані результати можна наступним чином:

1. Кластер за індексом 4, представляє високоякісні оголошення. Центральна точка характеризується високою ціною, яка становить близько 73976 доларів.
2. Кластер за індексом 3, представляє популярні та швидкі бюджетні оголошення. Тобто оголошення, близькі до цієї точки, будуть досить дешевими (близько 120 доларів) і матимуть низькі значення для необхідного мінімуму ночей (близько 3 ночей). Ці списки також мають багато відгуків (185) і тому є популярними місцями.
3. Кластер за індексом 2, представляє списки доступних місць для відпочинку. Ці оголошення є одними з найдешевших і становлять близько 111 доларів. Проте необхідна мінімальна кількість ночей найвища серед всіх кластерів, що може пояснюватись тривалою орендою за довгостроковим договором.
4. До кластера за індексом 1 входять споживачі, які витрачають близько 141 долар, що є дорожче, ніж користувачі кластерів з індексами 2 і 3. Він також має найнищу доступність (приблизно 31 день на рік). Отже, цей кластер представляє важкодоступні пропозиції середнього класу.
5. Кластер за індексом 0, має середнє значення для мінімальної кількості днів оренди близько 8, що є досить низьким показником, при цьому має найвищу доступність (278 днів). Отже, даний кластер є доступним для середнього класу.

Отже, проведена кластеризація виявила наступні ключові сегменти:

1. Доступні оголошення середнього діапазону (індекс 0)
2. Доступні довгострокові оголошення (індекс 1)
3. Оголошення тривалого проживання (індекс 2)
4. Популярні та бюджетні оголошення (індекс 3)
5. Оголошення високого діапазону (індекс 4)

Для проведення візуалізації отриманих результатів доцільно скористатись моделюванням за допомогою аналізу головних компонентів (PCA) — методу машинного навчання, який спрямований на зменшення розмірності наборів даних через зменшення масштабування бази даних [43].

Розділивши отримані кластери на 3 компоненти PCA, отримуємо наступний графік, що зображує визначені сегменти бази даних користувачів.

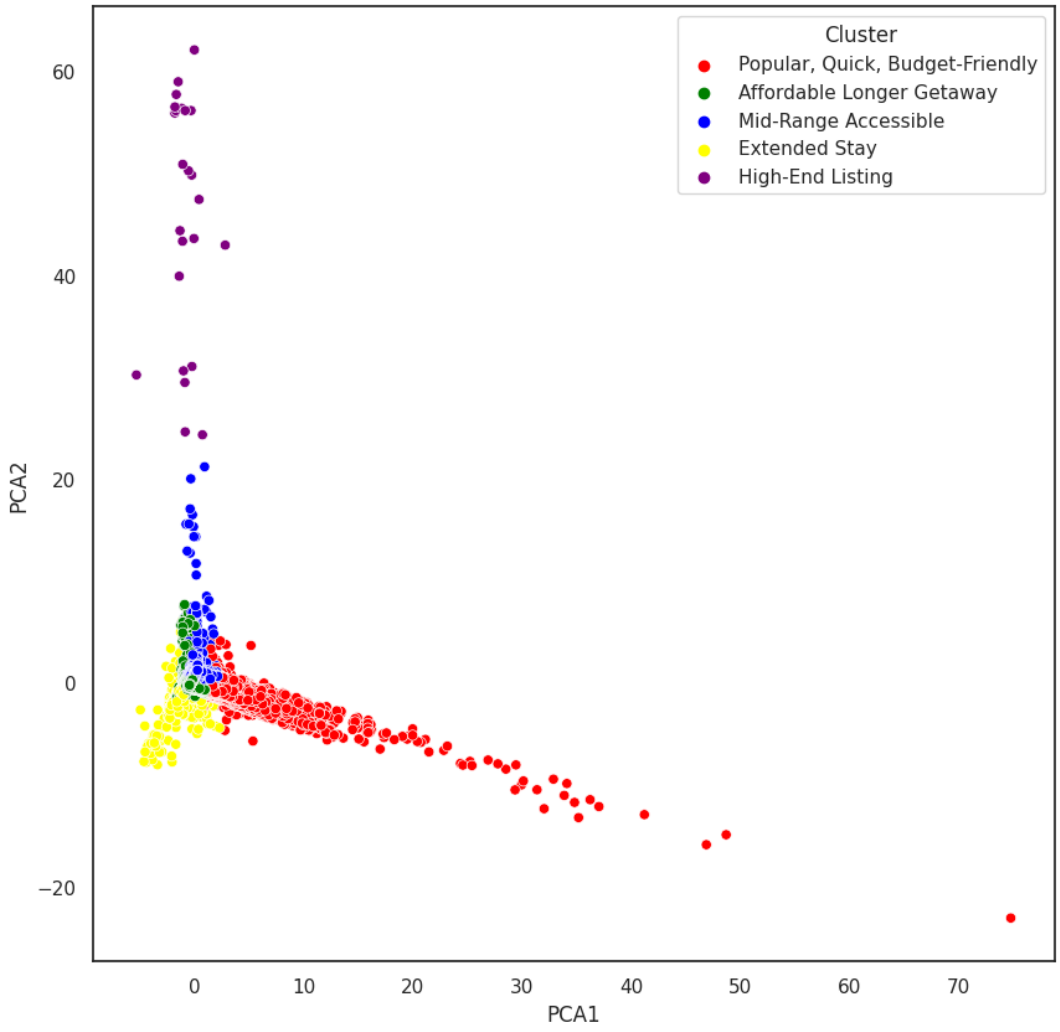

Рис. 3.9. Графік розподілу кластерів

Джерело: розрахунки автора

3.2.2. DBSCAN

Наступною моделлю класифікації є модель DBSCAN (Density-Based Spatial Clustering Application with Noise), яка побудована на базі 4 кластерів та має наступний вигляд:

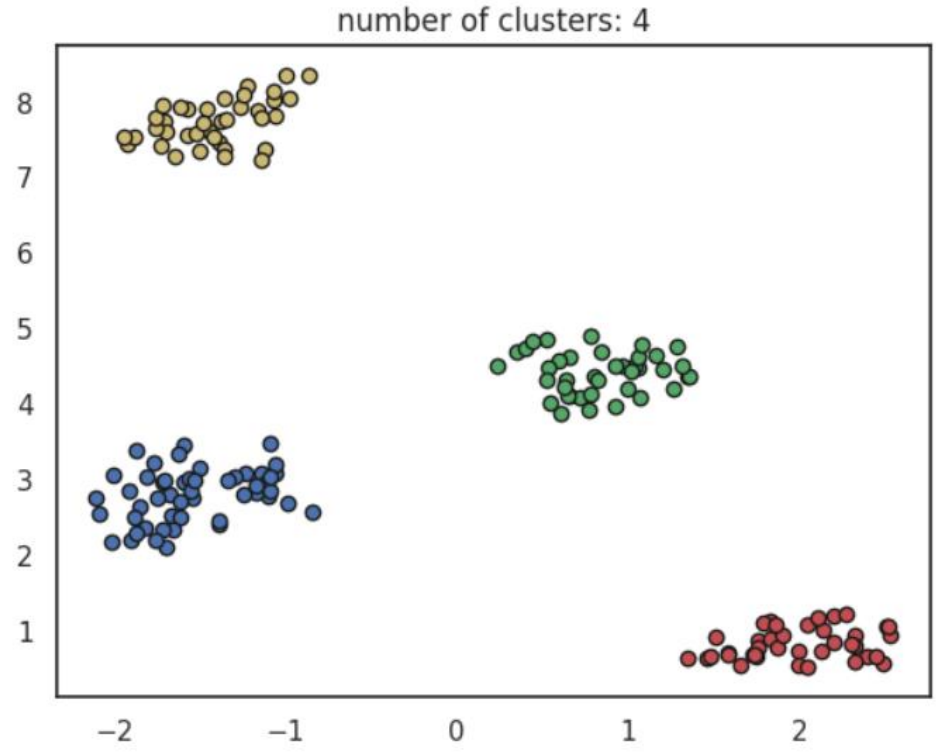

Рис. 3.10. Результат сегментації моделі DBSCAN

Джерело: розрахунки автора

Отже, при проведенні кластеризації за допомогою моделі DBSCAN можна виявити, що кластери досить добре відділені та мають високу чіткість, що свідчить про успішне застосування алгоритму.

3.2.4. BIRCH

Ще одною з моделей кластеризації є модель BIRCH (Balanced iterative reducing and clustering using hierarchies), результатом роботи якої є наступний графік:

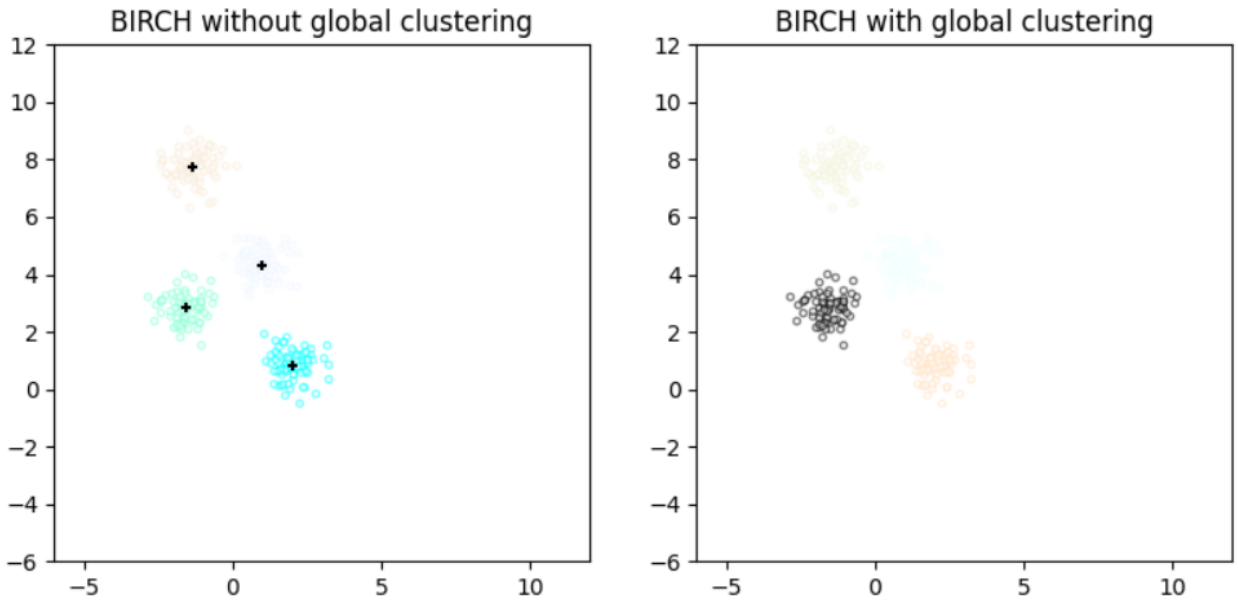

Рис. 3.11. Результат сегментації моделі BIRCH

Джерело: розрахунки автора

Можна проаналізувати, що кластери мають чітку межу та високу ефективність, що підтверджується визначенням ефективності моделі, результат якого показує, що для даної моделі розрахунок Adjusted Rand Index, що дорівнює значенню 1,00.

Зокрема, при оцінці якості моделей використовувалося декілька показників: Adjusted Rand Index, Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index. Реалізація даних індексів показує наступні результати:

Таблиця 3.1.

Порівняння ефективності моделей

Score	K-Means	DBSCAN	BIRCH
Adjusted Rand Index	0,91	0,31	1,00
Silhouette Coefficient	0,63	0,13	0,74
Calinski-Harabasz Index	1471,53	73,85	1742,14
Davies-Bouldin Index	0,66	2,49	0,37

Джерело: розрахунки автора

Якщо аналізувати якість представлених моделей, то найкращою моделлю, згідно з усіма реалізованими індексами, є модель BIRCH. З незначним відривом

наступною моделлю є K-Means, яка також показує відмінні результати при аналізі за допомогою Adjusted Rand Index, який дорівнює 0,91. Найгіршою моделлю серед досліджуваних виявилась модель DBSCAN. Проте, результати індексів для даної моделі є задовільними, що вказує на хорошу розподільчу здатність моделі до кластеризації.

Висновки до розділу 3

Отже, в даному розділі проведено аналіз бази даних, її очищення та візуалізація. В ході дослідження було реалізовано різні моделі кластеризації, серед яких: K-Means, DBSCAN, BIRCH.

При проведенні сегментації було застосовано кластеризацію на основі даних про оголошення, що містять інформацію про споживачів компанії Airbnb. Зокрема, було виявлено наступні групи, що базуються на клієнтській поведінці: доступні оголошення середнього діапазону, довгострокові оголошення, оголошення тривалого проживання, популярні та бюджетні оголошення та оголошення високого діапазону.

Результат порівняльного аналізу результативності моделей показав, що найкращою моделлю є BIRCH, другою по результативності – K-Means.

ВИСНОВОК

Дипломна робота має на меті проведення сегментації користувачів Airbnb на основі даних Airbnb для аналізу поведінки та потреб клієнтів, використовуючи декілька методів сегментації для визначення ключових характеристик та потреб різних сегментів клієнтів, а також проведення порівняльної характеристики реалізованих методів.

Отже, під час написання дипломної роботи було досліджено теоретичні аспекти впливу діяльності компанії Airbnb на ринок послуг з надання оренди, описано методичний підхід і інструментарій сегментування споживачів.

Дослідження розпочалось з огляду явища економіки спільного споживання, щоб окреслити її взаємозв'язок з галуззю туризму та розвитком концепції надання послуг з оренди. Наступним кроком було визначено місце та роль компанії Airbnb як прикладу цифрової платформи економіки спільного споживання. Зокрема, з'ясовано, що Airbnb займає друге місце у світі в галузі послуг оренди, що є підставою вважати дослідження корисним для розширення досліджень в цьому русі та виявленню нових можливостей для росту присутності компанії на ринку.

Також було визначено основні аспекти, за якими можна поділяти користувачів послуг компанії та проведено аналіз наявної літератури.

В другому розділі дипломної роботи були розглянуті методологічні засади сегментації, включаючи особливості та підходи до сегментації. Зокрема, було обрано проводити сегментацію на основі алгоритмів неконтрольованого машинного навчання із застосуванням моделей кластеризації. Серед моделей було обрано наступні: K-Means, DBSCAN та BIRCH.

Дослідження проводилось на основі даних, які були зібрані на сайті Inside Airbnb. Сформована база даних містила статистичну інформацію стосовно записів про оренду приміщення станом на квітень 2023 року, що публікувались користувачами Airbnb та розміщені у вільному доступі. База даних охоплює статистику 10 країн Європи.

Далі було проведено практичне застосування методів сегментації на реальних даних компанії Airbnb. Було проведено аналіз даних та реалізовані моделі кластеризації, зокрема K-Means, DBSCAN та BIRCH. Кожна з цих моделей була застосована до даних клієнтів Airbnb з метою отримання сегментації та виявлення груп клієнтів зі схожими характеристиками.

Одним з таких методів кластеризації є K-Means, що сегментував клієнтську базу на 5 груп, що засновуються на клієнтській статистичній інформації, яка зібрана зо допомогою вивчення оголошень компанії Airbnb. Виділені групи є наступними:

1. Доступні оголошення середнього діапазону (індекс 0);
2. Доступні довгострокові оголошення (індекс 1);
3. Оголошення тривалого проживання (індекс 2);
4. Популярні та бюджетні оголошення (індекс 3);
5. Оголошення високого діапазону (індекс 4).

Отримані результати аналізу даних та сегментації були представлені у вигляді таблиць і графіків. Результати кластеризації за допомогою моделей DBSCAN та BIRCH були також представлені у вигляді графіків та мають хороші показники кластеризації. Кожен кластер отримав позначення та був описаний у контексті його особливостей та значущості.

Порівняльний аналіз якості реалізованих моделей класифікації був проведений за допомогою наступних індексів: Adjusted Rand Index, Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index. В результаті обчислення було виявлено, що всі моделі мають задовільні показники та мають високу розподільчу здітність до класифікації. Проте, найкращою моделлю є BIRCH, а другою по результативності з невеликим відривом є K-Means.

В цілому, дослідження підтверджує, що використання методів сегментації на основі даних є корисним для розробки ефективних маркетингових стратегій в галузі гостьового розміщення. Це дослідження також дає можливість зрозуміти

поведінку та потреби клієнтів, що дозволяє компанії розробляти нові продукти та послуги, які відповідають їхнім потребам та очікуванням.

Використання аналітики клієнтів за допомогою сегментації на основі даних компанії Airbnb дозволило отримати важливі інсайти про поведінку та потреби клієнтів. Ці знання можуть бути використані для подальшого вдосконалення маркетингових стратегій, персоналізації послуг та збільшення задоволеності клієнтів.

Отже, використовуючи аналітику клієнтів за допомогою сегментації даних Airbnb, можна визначити різні сегменти клієнтів із чіткими вподобаннями, потребами та якостями. Цю інформацію можна використовувати для створення маркетингових кампаній, адаптованих до конкретних сегментів, що зрештою призведе до більшої задоволеності клієнтів.

Моделі для кластеризації, включаючи K-Means, DBSCAN і BIRCH, полегшують виявлення кластерів у наборі даних без вимоги попереднього визначення кількості кластерів. Це дозволяє виявити складні зв’язки та групи між клієнтами, які в іншому випадку залишилися б непоміченими традиційними методами аналізу даних.

Загалом, застосування клієнтської аналітики, яка базується на сегментації за допомогою даних Airbnb, є потужним інструментом для розуміння клієнтської бази, виділення значущих груп клієнтів і прийняття рішень щодо стратегій управління на основі обґрунтованих рішень.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Економіка спільної участі. Вікіпедія: веб-сайт. URL: https://uk.wikipedia.org/wiki/Економіка_спільної_участі
2. Zekanovic-Korona L., Grzunov J. Evaluation of shared digital economy adoption: Case of Airbnb. Electronics and Microelectronics (MIPRO). 2014. P. 1574-1579. URL: https://doi.org/10.1109/MIPRO.2014.6859816
3. Економічна модель спільного споживання (шерингова економіка). Дія.Бізнес: веб-сайт. URL: https://business.diia.gov.ua/handbook/impact-investment/ekonomicna-model-spilnogo-spozivanna-seringova-ekonomika
4. Hossain M. Sharing economy: a comprehensive literature review. International Journal of Hospitality Management. 2020. Vol. 87. URL: https://doi.org/10.1016/j.ijhm.2020.102470
5. Belk R. Why not share rather than own? The Annals of the American Academy of Political and Social Science. 2007. Vol. 611, № 1. P. 126–140. URL: https://doi.org/10.1177/0002716206298483
6. Price J.A. Sharing: the integration of intimate economies. Anthropologica. 1975. № 17. P. 3–27.
7. Belk R. You are what you can access: sharing and collaborative consumption online. Journal of Business Research. 2014. № 67. P. 1595–1600.
8. Munoz, P., Cohen, B. Mapping out the sharing economy: a configurational approach to sharing business modeling. Technological Forecasting and Social Change. 2017. № 125, P. 21–37. URL: https://doi.org/10.1016/j.techfore.2017.03.035
9. Economic impact reports. World Travel & Tourism Council 2022. URL: https://wttc.org/
10. Market size of the tourism sector worldwide from 2013 to 2022, with a forecast for 2023. Statista. 2023. URL: https://www.statista.com/statistics/1220218/tourism-industry-market-size-global/

11. Tourism set to return to pre-pandemic levels in some regions in 2023. World Tourism Organization. 2023. URL: https://www.unwto.org/news/tourism-set-to-return-to-pre-pandemic-levels-in-some-regions-in-2023
12. Wahab I. N. Role of information technology in tourism industry: impact and growth. International Journal of Innovative Research in Computer and Communication Engineering. 2007. № 4
13. 40+ fascinating Airbnb statistics (2023). Dream big, travel far. 2023. URL: https://www.dreambigtravelfarblog.com/blog/airbnb-statistics
14. Airbnb statistics [2023]: user & market growth data. Search logistics. 2023. URL: https://www.searchlogistics.com/learn/statistics/airbnb-statistics/
15. Airbnb statistics. IPropertyManagement. 2022. URL: https://ipropertymanagement.com/research/airbnb-statistics
16. Market capitalization of Airbnb (ABNB). CompaniesMarketcap. 2023. URL: https://companiesmarketcap.com/airbnb/marketcap/
17. NASDAQ: ABNB. Google Finance. 2023. URL: https://www.google.com/finance/quote/ABNB:NASDAQ?hl=uk&window=MAX
18. 68 Airbnb statistics, facts & figures. Awning. 2023. URL: https://awning.com/post/airbnb-statistics
19. Airbnb revenue and usage statistics (2023). Business of Apps. 2023. URL: https://www.businessofapps.com/data/airbnb-statistics/
20. Support for refugees fleeing Ukraine. Airbnb. 2022. URL: https://news.airbnb.com/help-ukraine/
21. Quattrone G., Kusek N., Capra L. A global-scale analysis of the sharing economy model – an AirBnB case study. EPJ Data Science. 2022. Vol. 11, № 36. URL: https://doi.org/10.1140/epjds/s13688-022-00349-3
22. Lutz C., Newlands G. Consumer segmentation within the sharing economy: The case of Airbnb. Journal of Business Research. 2018. № 88. P. 187–196. URL: https://iranarze.ir/wp-content/uploads/2019/01/E10648-IranArze.pdf

23. Vivek S. Clustering algorithms for customer segmentation. Towardsdatascience. 2018. URL: https://towardsdatascience.com/clustering-algorithms-for-customer-segmentation-af637c6830ac
24. Del Chiappa D., Sini G., L., Atzeni, M. A motivation-based segmentation of Italian Airbnb users: an exploratory mixed method approach. European Journal of Tourism Research. 2020. № 25. P. 2505. URL: https://www.researchgate.net/publication/358795264_A_motivation-based_segmentation_of_Italian_Airbnb_users_an_exploratory_mixed_method_approach
25. Anjana S. Customer segmentation using cluster analysis. Medium. 2020. URL:https://medium.com/analytics-vidhya/customer-segmentation-using-cluster-analysis-ed1f3a7c5920
26. How Airbnb uses data science to improve their product and marketing. Neil Patel. 2023. URL: https://neilpatel.com/blog/how-airbnb-uses-data-science/
27. Customer segmentation: How to segment customers? Kale. 2023. URL: https://blog.kale.bismart.com/en/customer-segmentation-how-to-segment-customers
28. Broom T., Chavez R., Wagner D. Becoming the king in the north: identification with fictional characters is associated with greater self–other neural overlap. Social Cognitive and Affective Neuroscience, 2021. Vol. 16, № 6, P. 541–551. URL: https://doi.org/10.1093/scan/nsab021
29. Felipe Ly J. Cluster analysis: market segmentation. LinkedIn. 2021. URL: https://www.linkedin.com/pulse/cluster-analysis-market-segmentation-juan-felipe-ly
30. What’s the difference between segmentation and clustering? Acquia. 2022. URL: https://www.acquia.com/blog/difference-between-segmentation-and-clustering

31. Customer segmentation via cluster analysis. Optimove. 2023. URL: https://www.optimove.com/resources/learning-center/customer-segmentation-via-cluster-analysis
32. K means clustering – introduction. GeeksforGeeks. 2023. URL: https://www.geeksforgeeks.org/k-means-clustering-introduction/
33. Yildirim S. K-Means clustering explained. Medium. 2020. URL: https://towardsdatascience.com/k-means-clustering-explained-4528df86a120
34. Dabbura I. K-means clustering: algorithm, applications, evaluation methods, and drawbacks. Medium. 2018. URL: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a
35. Difference between K-means and Hierarchical Clustering. GeeksforGeeks. 2023. URL: https://www.geeksforgeeks.org/difference-between-k-means-and-hierarchical-clustering/
36. Salton do Prado K. How DBSCAN works and why should we use it? Medium. 2017. URL: https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80
37. DBSCAN clustering in ML. Density based clustering. GeeksforGeeks. 2023. URL: https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/
38. BIRCH clustering. GeeksforGeeks. 2023. URL: https://www.geeksforgeeks.org/ml-birch-clustering/
39. Clustering. Scikit-learn. 2023. URL: https://scikit learn.org/stable/modules/clustering.html#birch
40. Inside Airbnb. URL: http://insideairbnb.com/about/
41. Reviewing inside airbnb: is it worth it? features, pricing, and reviews. Inside Airbnb. URL: https://airbtics.com/inside-airbnb-data/
42. Google Colab або Google Colaboratory: що це таке. Hardware Libre. URL: https://www.hwlibre.com/uk/google-colaboratory/

43. Whitfield B. A step-by-step explanation of principal component analysis (PCA). Built In. 2023. URL: https://builtin.com/data-science/step-step-explanation-principal-component-analysis

ДОДАТКИ

Додаток А

Реалізація практичної частини кваліфікаційної роботи

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import RobustScaler
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
import folium
df_Amsterdam = pd.read_csv("listings_Amsterdam.csv")
df_Amsterdam.head()
df_Athens = pd.read_csv("listings_Athens.csv")
df_Athens.head()
df_Barcelona = pd.read_csv("listings_Barcelona.csv")
df_Barcelona.head()
df_Berlin = pd.read_csv("listings_Berlin.csv")
df_Berlin.head()
df_Lisbon = pd.read_csv("listings_Lisbon.csv")
df_Lisbon.head()
df_London = pd.read_csv("listings_London.csv")
df_London.head()
df_Madrid = pd.read_csv("listings_Madrid.csv")
df_Madrid.head()
df_Paris = pd.read_csv("listings_Paris.csv")
df_Paris.head()
```

```
df_Rome = pd.read_csv("listings_Rome.csv")
df_Rome.head()
df_Vienna = pd.read_csv("listings_Vienna.csv")
df_Vienna.head()
# average for price
print("average price = ", df_Amsterdam["price"].mean())
print("average price = ", df_Athens["price"].mean())
print("average price = ", df_Barcelona["price"].mean())
print("average price = ", df_Berlin["price"].mean())
print("average price = ", df_Lisbon["price"].mean())
print("average price = ", df_London["price"].mean())
print("average price = ", df_Madrid["price"].mean())
print("average price = ", df_Paris["price"].mean())
print("average price = ", df_Rome["price"].mean())
print("average price = ", df_Vienna["price"].mean())
# average for minimum nights
print("average minimum nights = ", df_Amsterdam["minimum_nights"].mean())
print("average minimum nights = ", df_Athens["minimum_nights"].mean())
print("average minimum nights = ", df_Barcelona["minimum_nights"].mean())
print("average minimum nights = ", df_Berlin["minimum_nights"].mean())
print("average minimum nights = ", df_Lisbon["minimum_nights"].mean())
print("average minimum nights = ", df_London["minimum_nights"].mean())
print("average minimum nights = ", df_Madrid["minimum_nights"].mean())
print("average minimum nights = ", df_Paris["minimum_nights"].mean())
print("average minimum nights = ", df_Rome["minimum_nights"].mean())
print("average minimum nights = ", df_Vienna["minimum_nights"].mean())
# Summary statistics
df_Amsterdam.describe()
df_Athens.describe()
```

```
df_Barcelona.describe()
df_Berlin.describe()
df_Lisbon.describe()
df_London.describe()
df_Madrid.describe()
df_Paris.describe()
df_Rome.describe()
df_Vienna.describe()
# Most expensive listing
df_Amsterdam[df_Amsterdam["price"] == 7900.00]
# listings with 0 as the price
df_Amsterdam[df_Amsterdam["price"] == 0.00]
# listing with the highest value for minimum number of nights
df_Amsterdam[df_Amsterdam["minimum_nights"] == 1001.00]
# унікальні типи власності
# генеруємо список унікальних значень для 'room_type'
df_Amsterdam["room_type"].unique().tolist()
df_Athens["room_type"].unique().tolist()
df_Barcelona["room_type"].unique().tolist()
df_Berlin["room_type"].unique().tolist()
df_Lisbon["room_type"].unique().tolist()
df_London["room_type"].unique().tolist()
df_Madrid["room_type"].unique().tolist()
df_Paris["room_type"].unique().tolist()
df_Rome["room_type"].unique().tolist()
df_Vienna["room_type"].unique().tolist()
# знаходимо кількість унікальних значень для кожної країни
df_Amsterdam["room_type"].value_counts()
df_Athens["room_type"].value_counts()
```

```
df_Barcelona["room_type"].value_counts()
df_Berlin["room_type"].value_counts()
df_Lisbon["room_type"].value_counts()
df_London["room_type"].value_counts()
df_Madrid["room_type"].value_counts()
df_Paris["room_type"].value_counts()
df_Rome["room_type"].value_counts()
df_Vienna["room_type"].value_counts()
# Data visualizations
plt.figure(figsize = (10, 5))
sns.heatmap(df_Amsterdam.isnull(), yticklabels=False, cbar=False, cmap="Blues")
sns.heatmap(df_Athens.isnull(), yticklabels=False, cbar=False, cmap="Blues")
sns.heatmap(df_Barcelona.isnull(), yticklabels=False, cbar=False, cmap="Blues")
sns.heatmap(df_Berlin.isnull(), yticklabels=False, cbar=False, cmap="Blues")
sns.heatmap(df_Lisbon.isnull(), yticklabels=False, cbar=False, cmap="Blues")
sns.heatmap(df_London.isnull(), yticklabels=False, cbar=False, cmap="Blues")
sns.heatmap(df_Madrid.isnull(), yticklabels=False, cbar=False, cmap="Blues")
sns.heatmap(df_Paris.isnull(), yticklabels=False, cbar=False, cmap="Blues")
sns.heatmap(df_Rome.isnull(), yticklabels=False, cbar=False, cmap="Blues")
sns.heatmap(df_Vienna.isnull(), yticklabels=False, cbar=False, cmap="Blues")
capitals_lat_lng = {'Amsterdam': [52.377956,4.897070],
                    'Athens': [37.983810,23.727539],
                    'Barcelona': [41.390205, 2.154007],
                    'Berlin': [52.520008, 13.404954],
                    'Madrid': [40.403813, -3.741306],
                    'Lisbon': [38.736946, -9.142685],
                    'London': [51.509865, -0.118092],
                    'Paris': [48.864716, 2.349014],
                    'Rome': [41.902782, 12.496366],
```

```
                'Vienna': [48.210033, 16.363449]}

map = folium.Map(location=capitals_lat_lng['Madrid'],
                 tiles="Stamen Terrain", zoom_start=4.2)#Stamen Terrain
for key,val in capitals_lat_lng.items():
    folium.Marker(location=val).add_to(map)
map
# виведемо всі назви змінних
df_Amsterdam.columns
df_Athens.columns
df_Barcelona.columns
df_Berlin.columns
df_Lisbon.columns
df_London.columns
df_Madrid.columns
df_Paris.columns
df_Rome.columns
df_Vienna.columns
# використовуємо функцію concat() для створення нової бази даних "df_concat"
df_concat = pd.concat([df_Amsterdam_drop, df_Athens_drop, df_Barcelona_drop,
df_Berlin_drop, df_Lisbon_drop, df_London_drop, df_Madrid_drop, df_Paris_drop,
df_Rome_drop, df_Vienna_drop], ignore_index=True)
df_concat.head()
df_concat.tail()
df_concat.shape
# перевіримо чи значення колонки Area є унікальними та порахуємо їх кількість
df_concat["Area"].unique().tolist()
df_concat["Area"].value_counts()
# перевіримо чи наявні у базі даних відсутні значення
```

```
df_concat.isnull().sum()
# визначимо відсоток пропущених значень
percent_missing = df_concat.isnull().sum() * 100 / len(df_concat)
missing_values = pd.DataFrame({"percent_missing": percent_missing})
missing_values
# заповнюємо чарунки середнім значенням чисел
df_concat = df_concat.fillna(df_concat.mean())
# робимо перевірку
df_concat.isnull().sum()
# знайдемо рядки, де ціна = 0
df_concat[df_concat["price"] == 0]
# визначаємо середнє значення price
price_median = df_concat["price"].astype("float").median(axis = 0)
print("The median value of the price column: ", price_median)
# замінюємо знайденим значенням 0
df_concat["price"].replace(0, price_median, inplace = True)
# перевірка
df_concat[df_concat["price"] == 0]
df_concat.describe()
# Перевірка, чи немає повторюваних рядків у фреймі даних.
duplicate = df_concat.duplicated().sum()
print("Number of duplicate rows: ", duplicate)
# Форматування назв стовпців і стандартизація інформації
# очищення заголовків, видалення підкреслень та перетворення назв стовпців у формат заголовків
df_concat.rename(columns={"id": "ID", "latitude":"Latitude", "longitude":"Longitude", "neighbourhood":"Neighborhood", "room_type":"Room Type", "price":"Price", "minimum_nights":"Minimum Nights",
```

```
"number_of_reviews":"Number of Reviews", "reviews_per_month":"Reviews per Month", "availability_365": "Availability (365)"}, inplace=True)
df_concat.columns
df_concat.head()
# перевірка типів даних
df_concat.dtypes
df_concat["Price"] = df_concat["Price"].astype(float)
df_concat["Minimum Nights"] = df_concat["Minimum Nights"].astype(float)
df_concat["Number of Reviews"] = df_concat["Number of Reviews"].astype(float)
df_concat["Availability (365)"] = df_concat["Availability (365)"].astype(float)
df_concat.dtypes
df_concat.head()
sns.set_theme(style="white")
# Generate a large random dataset
d_matrix = df_concat
# Compute the correlation matrix
corr = d_matrix.corr()
# Generate a mask for the upper triangle
mask = np.triu(np.ones_like(corr, dtype=bool))
# Set up the matplotlib figure
f, ax = plt.subplots(figsize=(8, 7))
# Generate a custom diverging colormap
cmap = sns.diverging_palette(230, 20, as_cmap=True)
# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(corr, mask=mask, cmap=cmap, vmax=.3, center=0,
        square=True, linewidths=.5, cbar_kws={"shrink": .5})
# База даних містить категоріальні змінні. Щоб застосувати алгоритми машинного навчання, потрібно перетворити ці змінні в числовий формат (метод get_dummies())
```

```
dummies1 = pd.get_dummies(df_concat["Room Type"])
dummies1.head()
dummies1.rename(columns = {"Entire home/apt": "Entire Home", "Hotel room": "Hotel Room", "Private room": "Private Room", "Shared room":"Shared Room"}, inplace =True)
dummies1.head()
df_dummies1 = pd.concat([df_concat, dummies1], axis = 1)
df_dummies1.drop("Room Type", axis = 1, inplace = True)
df_dummies1.head()
dummies2 = pd.get_dummies(df_dummies1["Area"])
dummies2.head()
df = pd.concat([df_dummies1, dummies2], axis = 1)
df.drop("Area", axis = 1, inplace = True)
df.head()
df.tail()
# вибір потрібних змінних в базі даних та створення нової таблиці
df_sub = df[["Price", "Minimum Nights", "Number of Reviews", "Reviews per Month", "Availability (365)"]]
df_sub.head()
# нормалізація даних
scaler = StandardScaler()
df_scaled = scaler.fit_transform(df_sub)
df_scaled.shape
df_scaled
wcss = []
for i in range(1,11):
    kmeans_pca = KMeans(n_clusters = i, init = 'k-means++', random_state = 42)
    kmeans_pca.fit(df_scaled)
```

```
    wcss.append(kmeans_pca.inertia_)

# графік
plt.figure(figsize = (10,8))
plt.plot(range(1, 11), wcss, marker = 'o', linestyle = '-.',color='red')
plt.xlabel('Number of Clusters')
plt.ylabel('WCSS')
plt.title('K-means Clustering')
plt.show()
from sklearn.metrics import silhouette_score
# Тепер можна виконати кластеризацію K-середніх із 5 кластерами
# навчання моделі і створення об’єкту kmeans із 5 кластерами та передача йому даний, 'df_scaled'
kmeans = KMeans(n_clusters=5, random_state=42)
means = kmeans.fit(df_scaled)
# кластери, які пов'язані з кожним значенням даних
labels = means.labels_
# масив
means.cluster_centers_.shape
plt.scatter(df_scaled[:, 0], df_scaled[:, 1],
            c=means.labels_,
            s=70, cmap='Paired')
plt.scatter(means.cluster_centers_[:, 0],
            means.cluster_centers_[:, 1],
            marker='^', s=100, linewidth=2,
            c=[0, 1, 2, 3, 4])
# масштабування даних за допомогою зворотнього перетворення
cluster_centers = scaler.inverse_transform(cluster_centers)
cluster_centers = pd.DataFrame(data = cluster_centers, columns = [df_sub.columns])
```

```
cluster_centers
# Plotting the histograms associated with each cluster
for i in df.columns:
    plt.figure(figsize = (35, 5))
    for j in range(5):
        plt.subplot(1,5,j+1)
        cluster = df_cluster[df_cluster["Cluster"] == j]
        cluster[i].hist(bins = 20)
        plt.title("{}    \nCluster {} ".format(i,j))
      plt.show()
# Let's now concatenate the cluster labels to the 'pca_df' dataframe
pca_df = pd.concat([pca_df, pd.DataFrame({"Cluster": labels})], axis = 1)
# Display the first five rows
pca_df.head()
# Change the numeric values under the 'Cluster' column to the detailed cluster descriptions
pca_df["Cluster"].replace(0, "Mid-Range Accessible", inplace = True)
pca_df["Cluster"].replace(1, "Affordable Longer Getaway", inplace = True)
pca_df["Cluster"].replace(3, "Popular, Quick, Budget-Friendly", inplace = True)
pca_df["Cluster"].replace(4, "High-End Listing", inplace = True)
pca_df["Cluster"].replace(2, "Extended Stay", inplace = True)
# Display the first five rows
pca_df.head()
plt.figure(figsize=(10,10))
ax = sns.scatterplot(x="PCA1", y="PCA2", hue = "Cluster", data = pca_df, palette =["red","green","blue", "yellow","purple"])
plt.show()
# Display the correlations
df_cluster.corr()
```

```
correlations = df_cluster.corr()
f, ax = plt.subplots(figsize = (20, 10))
sns.heatmap(correlations, annot=True)
import matplotlib.pyplot as plt
import numpy as np
from sklearn.cluster import DBSCAN
from sklearn import metrics
from sklearn.datasets import make_blobs
from sklearn.preprocessing import StandardScaler
from sklearn import datasets
# Load data
df_sub, y_true = make_blobs(n_samples=300, centers=4,
                                    cluster_std=0.50, random_state=0)
db = DBSCAN(eps=0.3, min_samples=10).fit(df_sub)
core_samples_mask = np.zeros_like(db.labels_, dtype=bool)
core_samples_mask[db.core_sample_indices_] = True
labels = db.labels_
# Number of clusters in labels, ignoring noise if present.
n_clusters_ = len(set(labels)) - (1 if -1 in labels else 0)
print(labels)
# Plot result
# Black removed and is used for noise instead.
unique_labels = set(labels)
colors = ['y', 'b', 'g', 'r']
print(colors)
for k, col in zip(unique_labels, colors):
        if k == -1:
                # Black used for noise.
                col = 'k'
```

```
        class_member_mask = (labels == k)
        xy = df_sub[class_member_mask & core_samples_mask]
        plt.plot(xy[:, 0], xy[:, 1], 'o', markerfacecolor=col,
                        markeredgecolor='k',
                        markersize=6)
        xy = df_sub[class_member_mask & ~core_samples_mask]
        plt.plot(xy[:, 0], xy[:, 1], 'o', markerfacecolor=col,
                        markeredgecolor='k',
                        markersize=6)
plt.title('number of clusters: %d' % n_clusters_)
plt.show()
#evaluation metrics
sc = metrics.silhouette_score(df_sub, labels)
print("Silhouette Coefficient:%0.2f"%sc)
ari = metrics.adjusted_rand_score (y_true, labels)
print("Adjusted Rand Index: %0.2f"%ari)
import matplotlib.colors as colors
from sklearn.cluster import Birch, MiniBatchKMeans
from time import time
from itertools import cycle
# Use all colors that matplotlib provides by default.
colors_ = cycle(colors.cnames.keys())
fig = plt.figure(figsize=(12, 4))
fig.subplots_adjust(left=0.04, right=0.98, bottom=0.1, top=0.9)
# Compute clustering with BIRCH with and without the final clustering step
# and plot.
birch_models = [
    Birch(threshold=1.7, n_clusters=None),
    Birch(threshold=1.7, n_clusters=5),
```

```
]
birch_models

final_step = ["without global clustering", "with global clustering"]
for ind, (birch_model, info) in enumerate(zip(birch_models, final_step)):
    t = time()
    birch_model.fit(df_sub)
    print("BIRCH %s as the final step took %0.2f seconds" % (info, (time() - t)))
    # Plot result
    labels = birch_model.labels_
    centroids = birch_model.subcluster_centers_
    n_clusters = np.unique(labels).size
    print("n_clusters : %d" % n_clusters)
    ax = fig.add_subplot(1, 3, ind + 1)
    for this_centroid, k, col in zip(centroids, range(n_clusters), colors_):
        mask = labels == k
        ax.scatter(df_sub[mask, 0], df_sub[mask, 1], c="w", edgecolor=col, marker=".",
alpha=0.5)
        if birch_model.n_clusters is None:
            ax.scatter(this_centroid[0], this_centroid[1], marker="+", c="k", s=25)
    ax.set_ylim([-12, 12])
    ax.set_xlim([-12, 12])
    ax.set_autoscaley_on(False)
    ax.set_title("BIRCH %s" % info)
plt.show()
BIRCH_labels = labels
print('Silhouette Coefficient for BIRCH is:', metrics.silhouette_score(df_sub,
BIRCH_labels, metric='euclidean'))
ari = metrics.adjusted_rand_score (y_true, BIRCH_labels)
```

```
print("Adjusted Rand Index: %0.2f"%ari)
```