Міністерство освіти і науки України

Національний технічний університет

«Дніпровська політехніка»

(інститут)

інформаційних технологій

(факультет)

Кафедра системного аналізу та управління

(повна назва)

ПОЯСНЮВАЛЬНА ЗАПИСКА

КВАЛІФІКАЦІЙНОЇ РОБОТИ СТУПЕНЯ «БАКАЛАВР»

студента *Липко Софії Володимирівни*

(ПІБ)

академічної групи *124-20-2*

(шифр)

спеціальності *124 – Системний аналіз*

(код і назва спеціальності)

на тему: *«Візуалізація інформації щодо ефективності маркетингових кампаній за допомогою аналітичних інструментів Google»*

(назва за наказом ректора)

Керівники	Прізвище, ініціали	Оцінка за шкалою		Підпис
		рейтинговою	інституційною	
кваліфікаційної роботи	Доц. Одновол М. М.			
розділів:				
Інформаційно-аналітичний розділ	Доц. Одновол М. М.			
Спеціальний розділ	Доц. Одновол М. М.			
Рецензент	Доц. Кожевников А.В.			
Нормоконтролер	к.ф-м.н., доц. Хомяк Т.В.			

Дніпро

2024

ЗАТВЕРДЖЕНО:

завідувач кафедри

Системного аналізу та управління

(повна назва)

_____ *к. т. н., доц. Т.А. Желдак*

(підпис) (прізвище, ініціали)

« » 2024 р.

ЗАВДАННЯ

на кваліфікаційну роботу

ступеня бакалавра

(бакалавра, магістра)

студенту *Липко Софії Володимирівни* **академічної групи***124-20-2*

Спеціальності *124 - Системній аналіз*

на тему: *«Візуалізація інформації щодо ефективності маркетингових кампаній за допомогою аналітичних інструментів Google».*

затверджену наказом ректора НТУ «Дніпровська політехніка» *№469-с* від *23.05.2024 р.*

Розділ	**Зміст завдання**	**Термін виконання**
1.Інформаційно-аналітичний розділ	*Системний аналіз предметної області і постановка задачі.*	*10.11.2023-30.12.2023*
2. Спеціальний розділ	*Розробка моделі дашборду, схеми A/B тестування та розрахунок ефективності за MAI.*	*01.01.2024-05.06.2024*

Завдання видано _____ *доц. М. М. Одновол*

(підпис) (прізвище, ініціали)

Дата видачі: _____р.

Дата подання до екзаменаційної комісії 4 липня 2024 р.

Завдання прийняла до виконання _____ *Липко С. В.*

(підпис) (прізвище, ініціали)

РЕФЕРАТ

Кваліфікаційна робота містить 73 сторінок, 10 таблиць, 34 рисунки, 1 додаток. Список використаних джерел нараховує 12 найменувань.

Мета кваліфікаційної роботи – автоматизація аналізу основних показників відвідувань сайту по кільком маркетинговим кампаніям.

Об'єкт дослідження – процес розробки дашбордів з метою візуалізації KPI (ключових показників ефективності) та воронки продажу товарів.

Предмет дослідження – БД (база даних), створена системою GA4 (Google Analytics) та наповнена в процесі продажу товарів певного інтернет-магазину.

Методи дослідження: метод аналізу ієрархій та метод A/B тестуваня для малих груп. Обидва методи призначені для перевірки гіпотез та ухвалення рішень на основі зібраних даних. Також застосовується принцип об'єктно-орієнтованого підходу в Looker Studio.

Робота присвячена створенню візуальних аналітичних дашбордів і підтвердженню або спростовуванню гіпотез.

В інформаційно–аналітичному розділі наведено аналіз об'єкту дослідження та ключових проблем на ньому. Поставлені задачі дослідження та обрано концепції їх розв'язання.

У спеціальному розділі сформовано бізнес-логіку та алгоритми розрахунку ключових показників ефективності, розрахунки вибору гіпотези на основі методу аналізу ієрархій та методу A/B тестування.

Практична цінність отриманих результатів полягає в тому, що запропонований дашборд суттєво скорочує час обробки та аналізу великих обсягів даних та акцентує увагу на високих та недостатньо високих показниках ефективності.

***Ключові слова*:** база даних, аналіз даних, дашборд, показники ефективності кампанії, ефективність, візуалізація**.**

ABSTRACT

Master's degree work includes 73 pages, 10 tables, 34 drawings, 1 attachment. Bibliography has 12 items.

The purpose of the qualification work is analytical automatization of the main indicators of site visits for several marketing campaigns.

The object of the research is the process of developing tools for visualization of KPI (key performance indicators) and product sales funnel.

The subject of the research is a DB (database) created by the GA4 (Google Analytics) system and filled in the process of selling goods of the online store.

Research methods: hierarchy analysis method and A/B testing method for small groups. Both methods are designed to test hypotheses and make decisions based on the data collected. The principle of the object-oriented approach is also applied in Looker Studio.

The work is devoted to the creation of visual analytical dashboards and confirmation or refutation of hypotheses.

The information-analytical section provides an analysis of the object of research and key problems on it. The tasks of the study are set and the concepts of their solution are chosen.

In a special section, business logic and algorithms for calculating key performance indicators, hypothesis selection calculations based on the hierarchy analysis method and the A/B testing method are formed.

The practical value of the obtained results lies in the fact that the proposed dashboard significantly reduces the time for processing and analyzing large amounts of data and focuses on high and insufficiently high performance indicators.

Keywords: database, data analysis, dashboard, campaign performance indicators, effectiveness, visualization.

ЗМІСТ

ВСТУП 6

РОЗДІЛ 1. Дослідження предметної області і постановка задачі 7

1.1 Сутність та основні задачі створення аналітичних інформаційних панелей 7

1.2 Вибір показників ефективності маркетингових кампаній 13

1.3 Вимоги до якості та чистоти вхідних даних 17

1.4 Порівняння методу аналізу ієрархій та A/B тестування для прийняття рішень 21

1.5 Постановка задачі 43

РОЗДІЛ 2. Розробка моделі дашборду, схеми A/B тестування та розрахунок ефективності за MAI 44

2.1 Опис алгоритму отримання аналітичних метрик 44

2.2 Опис інтерфейсу користувача дашборду 51

2.3 Розрахунок ефективності кампаній за методом ієрархій 59

2.4 Підготовка та проведення A/B тестування 62

2.5 Оцінка результатів і побудова гіпотез підвищення ефективності кампаній 67

ВИСНОВКИ 71

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 72

ДОДАТКИ 73

ВСТУП

За останні десять років дані змінили обличчя нашого світу. Численні електронні листи, повідомлення в месенджерах, якими ми обмінюємося, відео на YouTube, фото в Instagram, є частиною майже 2,5 квінтильйонів байт даних, що генеруються щодня в усьому світі.

Компанії, як великі, так і малі, мають справу з величезними обсягами даних, і багато що залежить від їхньої здатності отримувати з них значущу інформацію. Аналітики даних роблять саме це – вони інтерпретують статистичні дані та перетворюють їх на корисну інформацію, яку компанії та організації можуть використовувати для прийняття важливих рішень.

Організації в усіх секторах все більше залежать від даних для прийняття важливих бізнес-рішень, наприклад, які продукти виробляти, на які ринки виходити, які інвестиції робити або на яких клієнтів орієнтуватися. Вони також використовують дані для виявлення слабких місць у бізнесі, які потребують вирішення.

Як результат, аналіз даних став однією з найбільш затребуваних IT-напрямків у всьому світі і, ймовірно, продовжуватиме зростати. Тому робота з величезними об’ємами даних та графічна візуалізація тенденцій є вкрай актуальною саме сьогодні.

Сучасні компаній стикаються з проблемою обробки великих обсягів даних і можуть втратити безліч часу на їх аналіз. На сьогодні головне не зібрати дані, а вчасно і якісно іх обробити. Однак, завдяки автоматизації процесу аналізу даних, компанії можуть швидше та ефективніше обробляти інформацію, що дає їм змогу швидше реагувати на зміни на ринку. Описова аналітика допомагає топ-менеджарам приймати правильні рішення щодо розробки нових товарів та поліпшення якості послуг, виявляти проблеми вже на етапі випуску на ринок нового застосунку і т. д.

Робота продуктового аналітика вкрай важлива, бо вона перетворює масиви незрозумілих даних в візуально зрозумілі інтерактивні інформаційні панелі (дашборди) з діаграмами, фільтрами, лініями тренду та висновками.

РОЗДІЛ 1. Дослідження предметної області і постановка задачі

1.1 Сутність та основні задачі створення статистичних інформаційних панелей

Бізнес-аналітика (Business Intelligence, скорочено BI) — це загальна назва для процесу збору, аналізу, інтерпретації та використання даних для забезпечення прийняття ефективних рішень у сфері бізнесу. Основна мета Business Intelligence полягає в тому, щоб зробити доступною для бізнес-користувачів інформацію, яку інакше було б складно отримати. Це сприяє кращому розумінню стану справ бізнесу, виявленню можливостей для вдосконалення та прийняттю обґрунтованих рішень.

Виділяють такі основні етапи BI процесу:

1. Отримання та збереження даних. Це можуть бути дані про встановлення нового застосунку і частоту користування ним.
2. Інтеграція та перевірка якості даних. На цьому етапі перевіряється цілісність даних, наявність дублікатів та прогалин в таблицях бази даних або електронної таблиці.
3. Аналіз даних. Це етап пошуку закономірностей в даних.
4. Візуалізація даних та інтерпретація висновків – останній етап, на якому власне і відбувається побудова діаграм, підрахунок основних аналітичних метрик і висвітлення тенденцій. На цьому етапі розробляється візуальний дашборд.

Дашборд (інформаційна панель) функціонально призначений для допомоги в прийнятті бізнес-рішень щодо вибору маркетингової стратегії по результатам тестових продажів товарів в інтернет-магазині, або по залученню геймерів в нову гру, або успішності зміни дизайну застосунку, тощо. Експлуатаційно така візуалізація потрібна для швидкої обробки даних, зібраних системою Google Analitics, Amplituda або іншою, та миттєвого виявлення тенденцій.

На сьогоднішній момент існує безліч комерційних і безкоштовних систем, здатних обробляти дані та візуалізувати показники конверсій, будувати воронки продажів і навіть формувати гіпотези на основі тенденцій. Серед таких інструментів можна виділити Tableau, Looker Studio, Power BI, Amplitude і навіть Google Sheets. Всі вони можуть працювати напряму з базою даних підприємства або можуть імпортувати таблиці даних в фоматах csv або json.

Щодо ергономічності, швидкості і візуальної досконалості, то лідером в цій галузі безумовно є Tableau, але це достатньо недешевий інструмент, а його безкоштовна версія Tableau Public не передбачає приватності, що не завжди доречно. Тобто, робочі книги не можна зберегти локально, але створені інформаційні панелі можна зберегти у загальнодоступному хмарному сховищі Tableau, до якого має доступ кожен. Разом з тим, Tableau може підключатися до простих баз даних (excel, pdf), до складних баз даних (Oracle), до баз даних у хмарі (веб сервіси Amazon), до баз даних Microsoft Azure SQL, Google Cloud SQL та інших джерел даних.

Основні галузі, де можна використати Tableau:

Комерційні підприємства. Багато компаній використовують Tableau для аналізу даних, створення звітів та візуалізації результатів. Відділи маркетингу, фінансів, продажів, логістики та управління проєктами можуть використовувати Tableau для отримання глибинних інсайтів із даних своєї компанії.

Громадський сектор. Урядові установи, неприбуткові організації та освітні заклади також використовують Tableau для аналізу даних та візуалізації результатів. Це може включати аналіз публічних даних, моніторинг програм та проєктів, а також створення звітів для прийняття рішень.

Медична галузь. Tableau використовується в медичних установах для аналізу клінічних даних, моніторингу результатів лікування та візуалізації

медичної статистики. Це дозволяє лікарям та адміністраторам здійснювати звітність, аналізувати тренди та приймати управлінські рішення на основі даних.

Фінансовий сектор. Банки, фінансові установи та страхові компанії використовують Tableau для аналізу фінансових даних, ризик-аналітики та створення звітів для внутрішнього та зовнішнього використання. Tableau допомагає виявляти тенденції, проводити порівняльний аналіз та прогнозувати фінансові результати.

Туризм та гостинності. Готелі, туристичні агентства та компанії, що надають готельні послуги, можуть використовувати Tableau для аналізу даних про бронювання, попиту на послуги, витрат та прибутку. Це допомагає виробляти ефективну стратегію управління та розуміти потреби клієнтів.

Сьогодні Tableau для аналітичної звітності використовують такі компанії, як LinkedIn, Amazon, Ferrari, Adobe, Cisco, Deloitte, Walmart та інші.

Інтерфейс Tableau Public (Рис. 1.1) доволі зручний і не потребує багато часу для освоєння:

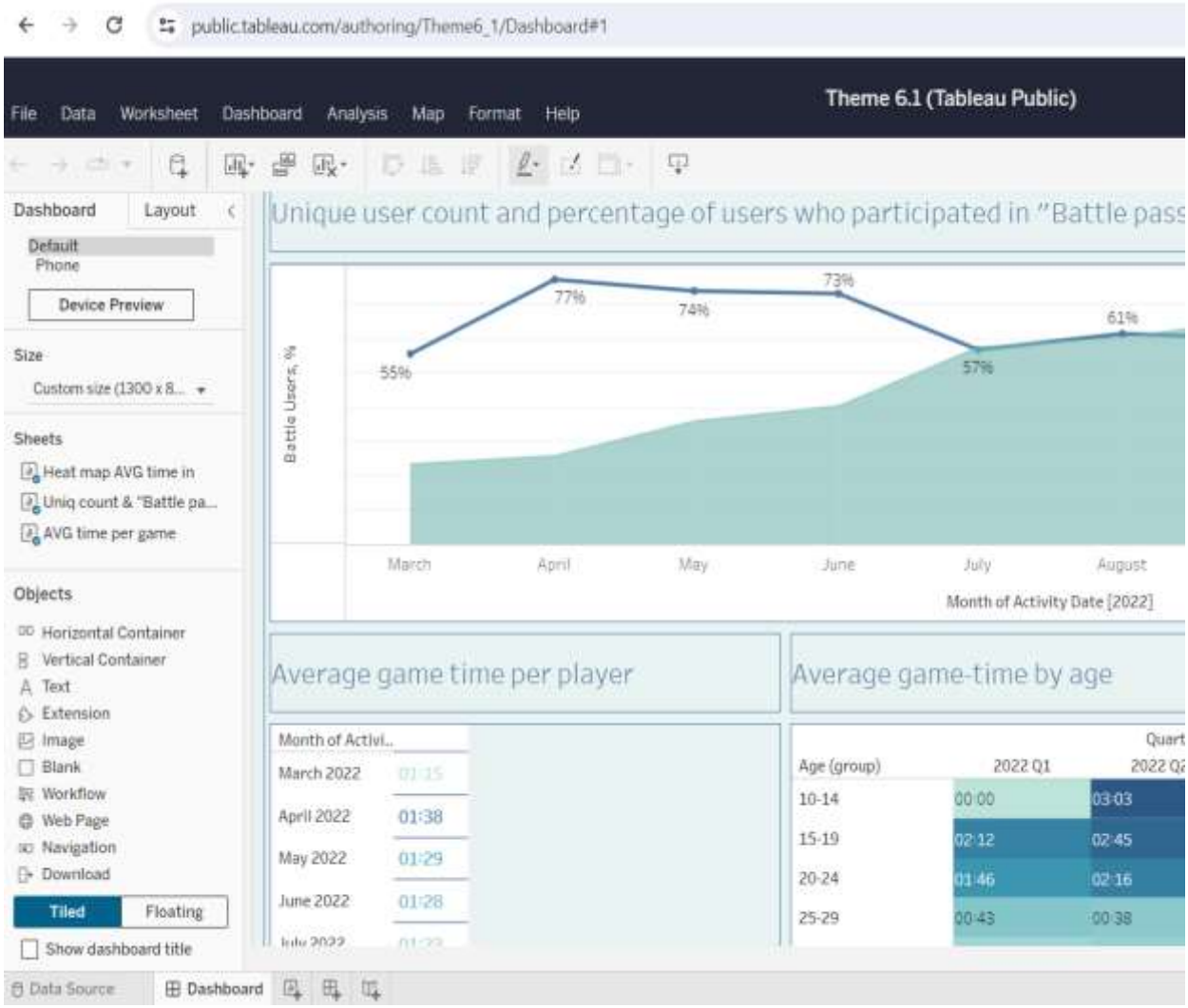


Рис. 1.1 Інтерфейс Tableau Public

Серед безкоштовних візуальних аналітичних інструментів лідером є Looker Studio від компанії Google (Рис. 1.2). Його перевага в тому, що він органічно взаємодіє з іншими інструментами Google, такими як BigQuery та Google Sheets (електронні таблиці онлайн). BigQuery - це хмарне сховище даних з повноцінною СУБД (системою управління базами даних), що дає можливість створювати SQL-запити до таблиць з даними, а також редагувати/додавати/видаляти самі таблиці (Рис. 1.3).

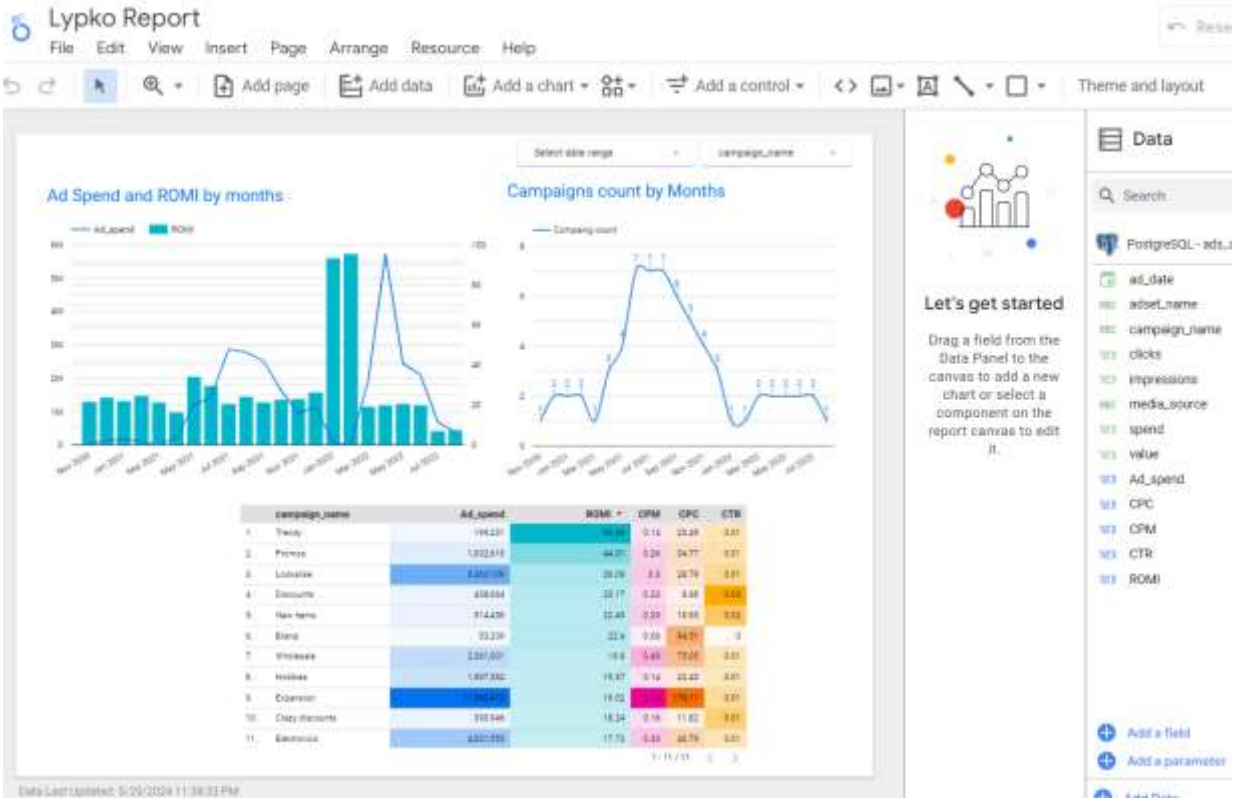


Рис. 1.2 Інтерфейс Looker Studio

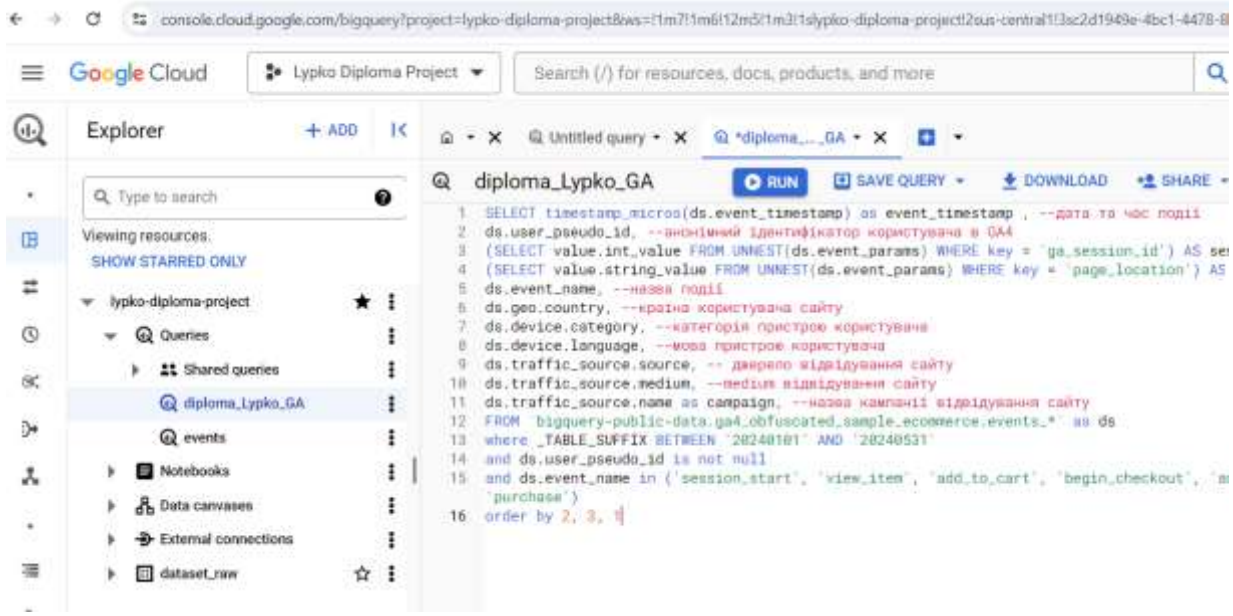


Рис. 1.3 Інтерфейс BigQuery

Хмарне сховище BigQuery дає можливість імпортувати дані в форматі *.csv з будь-якої іншої реляційної бази даних, наприклад локальної PostgreSQL, Oracle тощо. Якщо ж дані з сайту збирає GA4 (google analytics), то до такої схеми БД можна просто під’єднатись і зробити необхідну вібірку.

Потім цю вибірку використовують як Data Source (джерело даних). При бажанні вибірку можна зберегти як окрему таблицю (якщо запит складний і даних декілька мильонів рядків), це прискорить роботу дашборду. Якщо ж обсяг даних не такий великий, то є сенс працювати просто з SQL-запитом.

Інколи виникає необхідність працювати з власною локальною базою даних, тоді можна використати такий універсальний інструметн, як DBeaver. На Рис. 1.4 представлено всі СУБД, з якими працює цей інструмент.

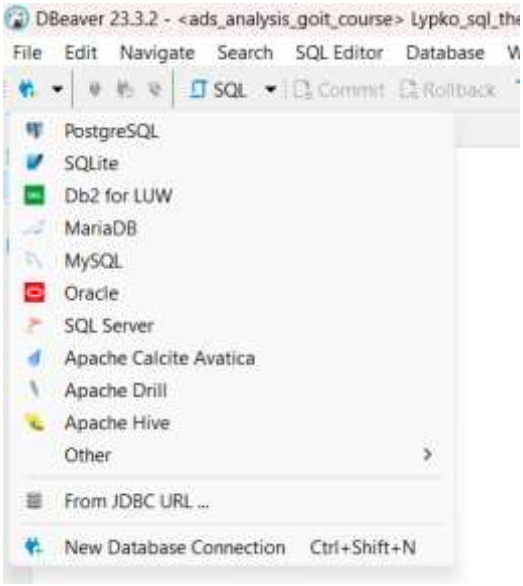


Рис. 1.4 СУБД, з якими працює DBeaver.

DBeaver допомагає доєднатись до бази даних, зробити вибірку лише необхіного, очистити дані від пустих значень та дублей, а потім експортувати дані в файл *.csv для подальшої роботи в BigQuery.

1.2 Вибір показників ефективності маркетингових кампаній

В сучасній Data-аналітиці існує велика кількість показників і зрізів даних. Серед них найбільш інформативні - це так звана «воронка» продажів і когортий аналіз.

Розберемо спочатку ключові показники:

Середнє арифметичне (Mean) — сума всіх значень, поділена на їхню кількість.

Усічене середнє арифметичне (Trimmed mean) — середнє значення після видалення фіксованої кількості екстремальних значень.

Зважене середнє арифметичне (Weighted mean) — сума всіх значень, помножена на вагу і поділена на суму ваг.

Медіана (Median) — міра, яка вказує на значення, що розділяє набір даних на дві рівні частини. Інакше кажучи, це значення, яке має половина даних вище нього, а половина — нижче.

Перцентиль (Percentile) — значення у розподілі, яке вказує на те, що певний відсоток даних знаходиться нижче цього значення. Наприклад, 25-ий перцентиль вказує на те, що 25% даних у розподілі мають значення менше або рівне цьому перцентилю, тоді як 75% даних мають значення більше або рівне цьому перцентилю. Медіана є 50-им перцентилем, оскільки вона розділяє набір даних на дві рівні частини.

Зважена медіана (Weighted median) — значення, яке вказує на те, що половина суми ваг лежить вище і нижче відсортованих даних.

Мода (Mode) — це значення даних або значення, які найчастіше зустрічаються у вибірці чи у наборі даних.

Екстремальне значення (Outlier) — значення даних, яке значно відрізняється від більшості даних.

Revenue є метрикою продукту, яка вимірює загальну суму грошей, отриману від продажу товарів або послуг.

Середній дохід на одного платника (ARPPU) — це кількість доходу що генерують платні користувачі, в середньому, за певний період часу.

Коефіцієнт конверсії (CR, Conversion Rate) — це ключова метрика продукту, що вимірює відсоток користувачів, які виконують бажану дію, таку як здійснення покупки, заповнення форми або завантаження застосунку. Це показник того, наскільки ефективний продукт у досягненні своїх бажаних результатів. Він може варіюватись залежно від продукту, бази користувачів та бажаної дії. Коефіцієнт конверсії може бути обчислений шляхом поділу загальної кількості користувачів, які виконали дію, на загальний розмір аудиторії, які переглянули оголошення, а потім множення цього числа на 100.

Відстеження та моніторинг коефіцієнта конверсії є важливим для оптимізації ефективності продукту. Існує кілька методів відстеження та моніторингу конверсійної ставки:

1. Встановити базову лінію. Перед тим, як почати оптимізацію коефіцієнта конверсії, важливо встановити базову лінію. Це надасть стартову точку для відстеження прогресу та виявлення областей для покращення. Базову лінію можна встановити, обчисливши коефіцієнт конверсії протягом певного періоду часу, наприклад, протягом місяця або кварталу.

2. Визначити «воронку» (Рис. 1.5). Для ефективного відстеження та моніторингу коефіцієнта конверсії важливо визначити воронку. Воронка — це послідовність кроків, які користувач повинен здійснити, щоб виконати бажану дію. Наприклад, воронка для електронної комерції може містити кроки, такі як перегляд товарів, додавання товарів до кошика та завершення процесу оформлення замовлення.

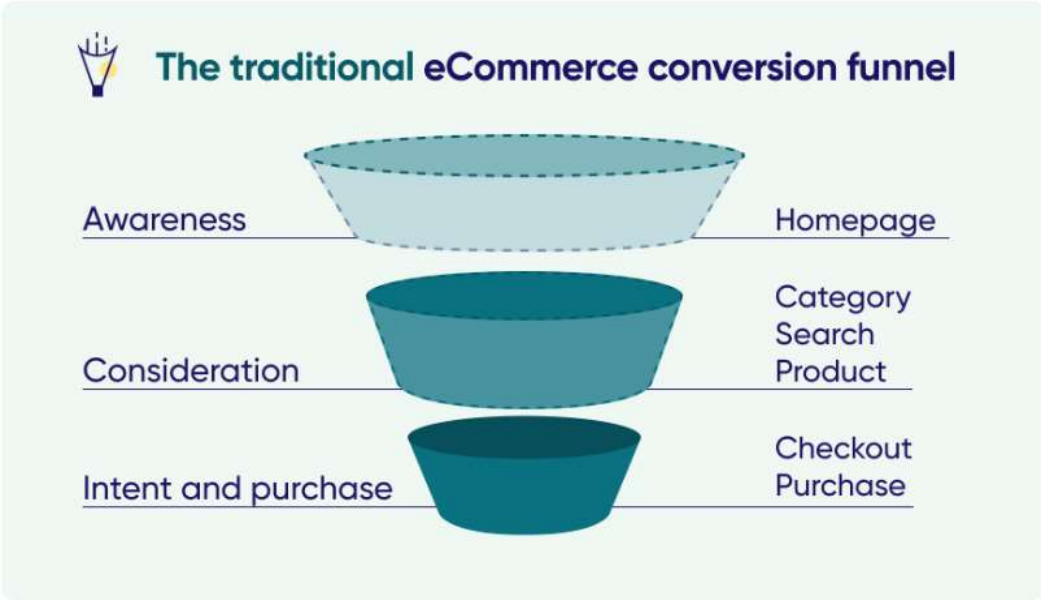


Рис. 1.5 Комерційна воронка

3. Аналіз поведінки користувачів. Аналіз поведінки користувачів може надати цінні висновки щодо коефіцієнту конверсії. Відстежуючи поведінку користувачів, аналітики даних можуть визначити, де користувачі “відпадають” у воронці та приймати рішення на основі даних для оптимізації користувацького досвіду. Наприклад, якщо користувачі “відпадають” під час процесу оформлення замовлення, аналітики даних можуть запропонувати спростити процес оформлення замовлення або запропонувати стимули для того, щоб користувачі завершили покупку.

4. Проведення A/B тестування. A/B тестування передбачає тестування різних версій продукту або маркетингової кампанії для визначення, яка версія генерує найвищий коефіцієнт конверсії. Порівнюючи коефіцієнти конверсії, згенеровані з різних версій, можна визначити яка з них є найефективнішою та відповідно вносити зміни.

Підходи аналізу для електронної комерції

Електронна комерція (e-commerce) — це процес купівлі та продажу товарів або послуг через Інтернет. Вона дозволяє здійснювати комерційні операції за допомогою онлайн-платформ, вебсайтів або мобільних застосунків.

Основні продуктові метрики:

Показник покинутих кошиків (CAR, shopping cart abandonment rate) вказує на відсоток користувачів, які додали товар до кошика, але так і не зробили замовлення. Більшість з них насправді були готові здійснити покупку, але в останній момент відволіклися або передумали. Для розрахунку цього показника використовуються дані з вебаналітики.

Середній чек (AOV, average order value) є найпоширенішим показником і однією з надзвичайно важливих метрик у довгостроковій перспективі. Він розраховується як відношення загального доходу, отриманого від продажів або виконання замовлень, до загальної кількості цих замовлень протягом певного періоду часу. За допомогою AOV можна прогнозувати дохід інтернет-магазину, враховуючи конверсію і кількість користувачів, а також коригувати стратегію розвитку. Якщо середній чек низький, то може бути вигідніше стимулювати його збільшення, а не приваблювати нових покупців.

Показник прибутковості клієнта (CRR, cost revenue ratio) відображає співвідношення між тим, скільки витрачено на рекламу і доходом від продажів. Чим нижче CRR, тим успішніший бізнес.

Окупність витрат на рекламу (ROAS, return of advertising spent) — окупність витрат на рекламу. Це зворотній до CRR показник. Він повинен бути понад 100%.

1.3 Вимоги до якості та чистоти вхідних даних

Кожен аналітик знає, що робота з якісними даними відіграє ключову роль в отриманні правдивого кінцевого результату. Аналіз даних — це послідовність кроків, які необхідно виконати, щоб зрозуміти та логічно осмислити наявні дані. Зазвичай легко визначити, який етап аналізу даних є найважливішим, але не слід забувати, що всі етапи є важливими: вони дозволяють впевнитись, що ми правильно оцінюємо дані, що результати є корисними та можуть бути застосовані на практиці.

Рис. 1.6 Основні етапи аналізу даних

Основні етапи аналізу даних:

1. Визначення завдання. Спершу треба зрозуміти, для чого проводиться аналіз, які дані потрібні та що саме варто аналізувати.

2. Збір даних та перевірка якості. Після формування завдання дані збираються з джерел, таких як опитування, інтерв'ю, анкети, пряме спостереження та фокус-групи, трекінг поведінки користувача на сайті або в застосунку, тощо. Важливо організувати зібрані дані для подальшого аналізу та перевірити їх якість.

3. Очищення та організація даних. Отже, не всі зібрані дані будуть корисними, тому на цьому етапі потрібно їх очистити. У цьому процесі видаляються прогалини, помилки та дублікати. Слід підкреслити, що очищення даних є обов’язковим етапом аналізу. Крім того, організація даних для покращення їх візуалізації є надзвичайно важливим аспектом для прийняття ефективних бізнес-рішень, бо якщо аналітик не зможе побачити

всі свої важливі дані в одному місці та зрозуміти, як вони пов'язані між собою, то він стикатиметься з труднощами у прийнятті обґрунтованих рішень.

4. Аналіз даних. На цьому етапі вже використовується програмне забезпечення для аналізу даних та інші інструменти з метою подальшої їх інтерпретації та формування висновків. Інструменти аналізу даних включають Google Sheets, SQL, BigQuery, Python, Power BI, Google Looker Studio та інші.

5. Інтерпретація та візуалізація. На фінальному етапі необхідно інтерпретувати дані та обрати найкращий шлях для вирішення досліджуваного завдання. Візуалізація даних допомагає графічно відобразити інформацію так, щоб зацікавлені сторони могли її читати та розуміти. Можна використовувати діаграми, графіки, теплові карти, маркери чи багато інших методів. Візуалізація допомагає отримувати цінні ідеї, допомагаючи порівнювати набори даних та спостерігати зв'язки між ними.

Зупинимось більш детально на етапі збору даних. Джерела даних можна поділити на п’ять категорій:

- внутрішні бази даних та технічні документи компанії (наприклад, БД сайту).
- дані з сервісів, які використовує компанія. Більшість 3rd party services дозволяє забирати звідти дані автоматично, наприклад, через API. Приклади сервісів: CRM, ERP, Ads (Google, Facebook), Email marketing тощо.
- ПЗ для аналітики та відстежування поведінки користувачів: GA4, Amplitude, Heap analytics.
- таблиці Excel та Google sheets, які завжди є майже в будь-якій компанії.

Якість даних відіграє важливу роль в аналізі даних, оскільки будь-які неточності або помилки в даних можуть призвести до неправильних

висновків та рішень. Крім того, некоректні дані можуть призвести до непотрібних витрат на додаткові дослідження та виправлення помилок.

Перед початком роботи з даними важливо переконатися, що вони відповідають критеріям якості:

- Точність (Accuracy) — дані є точними. Наприклад, дані про оплати не округлюються, а відображаються з точністю до копійки.
- Повнота (Completeness) — дані є повними. Наприклад, всі дані про старт сессії в користувачів сайту присутні.
- Консистентність (Consistency) — дані є логічно пов'язаними між собою та взаємодіють без протиріччя. Наприклад, всі дати продажів зберігаються в єдиному форматі дати DD.MM.YYYY.
- Своєчасність (Timeliness) — ми отримуємо дані вчасно для того, щоб зробити висновки. Наприклад, дані про продажі через місяць після проведення рекламної кампанії не допоможуть оптимізувати рекламу.
- Валідність (Validity) — дані є валідними в бізнес-контексті. Наприклад, дані про продажі містять тільки найменування товарів, які присутні в локальній CRM.
- Унікальність (Uniqueness) — дані не дублюються. Наприклад, одна дія клієнта не записується кілька разів.

В підсумку можна сказати, що для ефективного аналізу даних необхідно належним чином підготувати та перевірити якість доступних даних. Джерела даних можуть бути різноманітними, від внутрішніх баз даних компанії до зовнішніх сервісів та соціальних мереж, однак перед початком роботи з даними, необхідно переконатися в їх точності, повноті, узгодженості, своєчасності, валідності та унікальності. Крім того, якість даних може оцінюватися суб'єктивно, особливо в складних системах.

Тепер варто зупинитись на етапі очищення даних. Data cleaning (очистка даних) — це спосіб підвищення якості отриманих даних.

Тут можна зробити наступне:

- прибрати дублікати;
- зробити дані консистентними у своєму форматі;
- зробити дані унікальними;
- прибрати невалідні записи.

Можна навіть покращити повноту даних (completeness) через певний метод інтерполяції (interpolation).

Існує багато інструментів для очищення даних і для подальшого аналізу. Ось деякі з найпопулярніших:

- Google Sheets — один з інструментів, який найбільш широко використовується для аналізу даних. Google Sheets дозволяє проводити базовий аналіз даних, зокрема сортування, фільтрацію та групування даних.
- SQL — мова програмування, яка використовується для роботи з базами даних. За допомогою SQL можна вилучати дані з БД, змінювати їх, а також проводити різні операції, такі як сортування та групування даних. SQL широко використовується в галузях, де важливо обробляти великі обсяги даних.
- Python — мова програмування, яка використовується в різних галузях, зокрема в таких як наука про дані та штучний інтелект. Python надає багато бібліотек та модулів для обробки даних та аналізу, такі як Pandas, NumPy та Matplotlib.
- Tableau або Power BI — онлайн інструменти для візуалізації даних, які дозволяють створювати інтерактивні діаграми та графіки, а також створювати звіти та інформаційні панелі. Часто ці інструменти використовують в комплексі.

1.4 Порівняння методу аналізу ієрархій та A/B тестування для прийняття рішень

В умовах сьогодення продуктовому аналітику потрібно не тільки зібрати дані, очистити їх, розробити інформаційну панель метрик, а й провести оцінку ефективності впровадження рекламних кампаній на тестових групах користувачів. Для коректного прогнозування успішності великої маркетингової кампанії важливо мати математичний алгоритм, здатний проаналізувати весь набір важливх критеріїв і запронувати висновок, який не буде залежати від суб'єктивного впливу аналітика. В усьому світі для комерційних маркетингових кампаній використовується простий і зрозумілий алгоритм A/B тестування на контрольних групах, але можна спробувати використати класичний метод сиситемного аналізу – метод ієрархій.

Метод аналізу ієрархій (MAI) дуже поширений метод для рішення багатокритеріальних задач. MAI − це математичний інструмент системного підходу до складних проблем прийняття рішень, запропонований американським математиком Томасом Сааті. Процес прийняття рішень поєднує в собі психологічні аспекти і математичні методи, тому що альтернативи отримують експертні оцінки у відповідності до розуміння проблеми дослідниками, але далі інформація, що отримана, структурується та аналізується за допомогою математичного метода. Цей метод не надає людині, яка приймає рішення, якоїсь чіткої, однозначно правильної відповіді, а дозволяє їй самостійно знайти таку альтернативу, яка б максимально корелювалася із її власним розумінням суті проблеми та найкращого шляху її вирішення. MAI уможливлює відносно просту і обґрунтовану структуризацію проблеми прийняття рішень у вигляді ієрархії, отримавши на виході кількісні показники – оцінку кожного варіанту вирішення поставленої проблеми, чи в термінах методу - альтернативи.

Найпростіша форма, застосовувана для структурування проблеми – це ієрархія, що складається з трьох рівнів: цілі процесу прийняття рішення – на першому рівні, критерії за якими альтернативні варіанти досягнення цієї цілі, що розташовані на третьому рівні, будуть зважені.

Насправді, ієрархічна декомпозиція складних систем – це метод, яким людський мозок за довгі роки еволюції звик вирішувати складні проблеми. Йому притаманно організовувати фактори впливу на рішення у порівневому порядку починаючи з найважливішого, що розташовується на найвищому рівні ієрархії, до менш важливих, які знаходяться, відповідно на нижчих її рівнях. Ціллю подібного структурування є уможливлення здійснення судження щодо важливості елементів на визначеному рівні з оглядом на елементи суміжних рівнів, що дозволяє максимально впорядкувати складові проблеми і швидше зрозуміти її суть. Достатнім слід вважати той набір критеріїв та у тій кількості, застосувавши який проблему можна буде повноцінно вирішити. Тому, одна з основних задач експерта – знайти той нестійкий баланс між необхідним та надлишковим.

Порядок застосування методу аналізу ієрархій: процес починається першим етапом, на якому здійснюється побудова повноцінної моделі вирішуваної проблеми у вигляді ієрархії. До ієрархії прийнято долучати ціль, яка переслідується, критерії, на основі яких рішення проблеми буде знайдено та альтернативи, серед яких буде здійснено вибір. Приклад представлено на Рис 1.7.

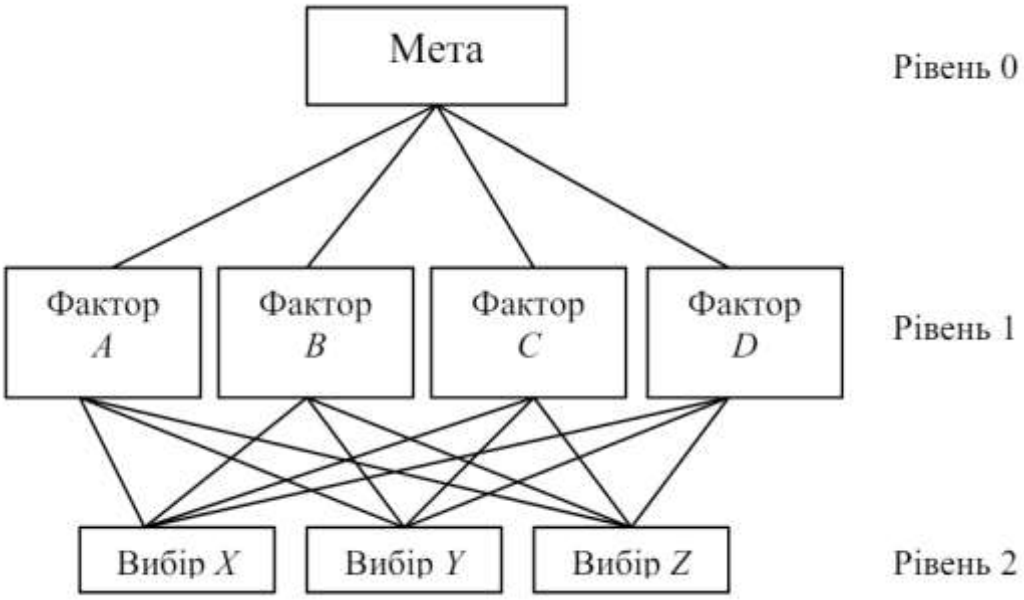


Рис. 1.7 Найпростіша архітектура в методі аналізу ієрархій

Метод аналізу ієрархій включає наступні основні етапи:

1 етап: декомпозиція проблеми;

2 етап: побудова ієрархічної структури моделі проблеми;

3 етап: експертне оцінювання переваг, де будуються матриці парних порівнянь (спочатку порівнюються один з одним всі критерії в контексті їх важливості для розв'язання задачі, потім для кожного з критеріїв виконується парне порівняння можливих рішень з точки зору оцінки їх важливості в межах даного критерію. В результаті отримаємо одну матрицю [$nnnnn$] для порівняння критеріїв і n матриць [$mmmmmm$] для порівняння альтернатив у межах кожного критерію;

4 етап: побудова локальних пріоритетів LL. В кожній з матриць парних порівнянь RR виконуються наступні дії: будують нормалізовану матрицю NN (спочатку знаходять суму єлєментів кожного стовпця матриці парних порівнянь RR, потім ділять кожен елемент стовпця на відповідну суму) і обчислюють середнє арифметичне (геометричне) рядків нормалізованої матриці парних порівнянь;

5 етап: оцінка узгодженості матриць парних порівнянь. Неузгодженість матриць парних порівнянь полягає в порушенні транзитивності відношення переваги (якщо A > B і B > C, то $\Longrightarrow$ A > C); існує алгоритм, наведений нижче, який дозволяє визначити оцінку погодження;

6 етап: синтез глобальних пріоритетів, необхідно помножити матрицю локальних пріоритетів, які відповідають рішенням, на вектор – рядок пріоритетів, які відповідають критеріям. Отримаємо вектор – стовпець пріоритетів, які відповідають рішенням;

7 етап: висновки й пропозиції для прийняття рішень;

8 етап: формулювання рішення.

Метод аналізу ієрархій при побудові єдиної шкали для різних компонентів проблеми використовує міру ступеня впливу кожного фактору одного рівня на фактори верхнього рівня або на кінцеву мету. Ця міра утвориться в результаті висловлення суджень про ступінь впливу (важливості) цих факторів.

Парне порівняння (ПП) елементів рівня виконується за результатом знань, суджень, досвіду, преференцій та думок усіх акторів процесу прийняття рішення, але чисельні переваги цих елементів визначаються згідно ранжуванню фундаментальної шкали значень (Таблиця 1.1), з якої можна побачити різновиди кількісної та відповідної якісної оцінки переваги одного елемента над іншим.

Таблиця 1.1 Фундаментальна шкала значень в методі аналізу ієрархій

Судження	Тлумачення
1. Рівнозначна важливість	Рівнозначна важливість факторів по відношенню до цілі
2. … Проміжне судження	
3. Помірна перевага	Досвід та судження надають невелику перевагу одного фактору над іншим

4. … Проміжне судження	
5. Відчутна перевага	Відчутна перевага одного фактора над іншим
6. … Проміжне судження	
7. Суттєва перевага	Майже тотальна перевага одного фактора над іншим
8. … Проміжне судження	
9. Тотальна перевага	Тотальна перевага одного фактора над іншим
1/k	Обернені величини

Для порівняння n об’єктів складають матрицю парних порівнянь R [$nnnnnn$] (завжди квадратна):

Таблиця 1.2 матриця парних порівнянь

Об’єкти	O1	O2	…	On
O1	1	A12	…	A1n
O2	A21	1	…	A2n
…	…	…	1	…
On	An1	An2	…	1

Елемент матриці Aij – міра переваги об’єкта Oi над об’єктом Oj, яка виражається експертом у шкалі Сааті й приймає значення від 1 до 9. Діагональні елементи матриці завжди дорівнюють 1. У матриці ПП симетричні елементи логічно відповідають правилу:

$$Aij = \frac{1}{Aji}$$

Експерт заповнює тільки верхню наддіагональну частину матриці ПП і при цьому кількість необхідних суджень експерта обмежується числом: $\frac{n(n-1)}{2}$, де n – кількість порівнювальних об’єктів.

Розрізняють шкали:

- простого порядку (шкала найменувань);
- слабого порядку (рангова шкала, шкала Сааті);
- сильного порядку (шкала інтервалів, шкала відношень, абсолютна шкала).

Шкала простого порядку складається тільки з списку об’єктів з вказанням порядкового номера; об’єкти не порівнюються. Шкала слабого порядку: виконується порівняння об’єктів в змісту ‘‘гірше, краще, такий самий”; обчислюються ранги (упорядкування). Шкала сильного порядку: не тільки виконується упорядкування об’єктів за будь-якою властивістю, а й чисельно визначаються сила (степінь) цієї переваги.

Розрізняють дві ситуації експертних оцінок порівняльної важливості об’єктів:

Ситуація 1 має місце коли міра порівнювальних властивостей виражена в сильних шкалах (для упорядкування об’єктів чисельно визначається їх міра властивості; шкала інтервалів, шкала відносин, абсолютна шк.). У цьому випадку, якщо міра властивості об’єкта $Oi = wi$, а міра об’єкта $Oj = wj$, то міра переваги Oi порівняно з Oj:

$$\frac{Oi}{Oj} = \frac{wi}{wj}$$

Матриця переваг у цьому випадку є узгодженою. Узгодженість означає, що якщо порівнюються n об’єктів, то достатньо $(n - 1)$ думок щодо їх порівняння.

Ситуація 2 полягає в тому, що властивості об’єктів слабо структуровані і можуть бути оцінені тільки в шкалі слабого порядку. Виконується порівняння об’єктів у змісті ‘‘гірше, краще, такий самий’’, для упорядкування обчислюються ранги (рангові шкали, шкали Сааті). В такому випадку, при використанні шкали Сааті, неможливо досягти повної

узгодженості через різні кваліметричні шкали у різних об’єктів тому порушується питання про ступінь погодженості отриманих оцінок.

Як міру погодженості розглядають два показники:

1. Індекс узгодженості (ІУ);

2. Відношення узгодженості (ВУ).

В загальному випадку узгодженість зворотньосиметричної матриці еквівалентна до вимоги рівності її максимального власного значення λ max числу порівнювальних об’єктів n:

$$\lambda \text{max} = n$$

В якості міри узгодженості розглядають нормоване відхилення λmax від n, таку міру називають індексом узгодженості ІУ:

$$\text{ІУ} = \frac{(\lambda max - n)}{(n-1)}$$

Оцінка прийнятності отриманого узгодження виконується порівнянням його з випадковим індексом (ВІ). Випадковий індекс – це індекс узгодженості, розрахований для квадратної n–вимірної додатної зворотньосиметричної матриці, елементи якої згенерували датчиком випадкових чисел, розподілених за рівномірним законом для інтервалу значень: 1/9, 1/8, 1/7, 1/6, 1/5, 1/4, 1/3, 1/2, 1, 2, 3, 4, 5, 6, 7, 8, 9.

У таблиці 1.3 представлені значення випадкового індексу ВІ для різних матриць порядку від 2 до 15.

Таблиця 1.3

Порядок матриці n	1	2	3	4	5	6	7	8	9
ВІ	0	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49

Порядок матриці n	10	11	12	13	14	15
ВІ	1.51	1.54	1.56	1.57	1.59	1.24

Відношення узгодженості ВУ обчислюється за формулою:

$$\text{ВУ} = \frac{\text{ІУ}}{\text{ВІ}}$$

Якщо ВУ $\leq$ 0,1 або ВУ$\in$ (0,1; 0,3), то рівень узгодженості є прийнятним.

Якщо ВУ > (0,1; 0,3), то відношення узгодженості – неприйнятне, в такому випадку, експерту рекомендується переглянути свої думки.

Аналіз результатів експертних оцінок полягає в отриманні вектору пріоритетів порівнювальних об’єктів.

A/B тестування – сучасний інструмент ананлітика для дослідження та прогнозування результатів впровадження кампаній. A/B-тестування (інша назва - спліт-тести) — це підхід, який допомагає командам перевіряти гіпотези та ухвалювати рішення на основі даних, а не інтуїції. Суть методу полягає в тому, щоб розділити аудиторію на частини та показати їй різні варіанти реклами, щоб зрозуміти, яка версія краща.

Тестувати можна дизайн застосунка, колір та розташування ключової кнопки, форму реєстрації, оформлення email-розсилки, текст оголошення на сайті та інші зміни.

Результати A/B-тесту показують, яке рішення дасть більшу конверсію в потрібну цільову дію. Наприклад, у якому разі більше користувачів перейде за посиланням, зареєструється на сайті або в додатку, підпишеться на розсилку, заповнить форму зворотного зв’язку. Є і складніші тести, спрямовані дослідження довгострокових метрик, як-от середній чек чи вплив змін у продукті з прибутку.

В основі A/B тестування лежить математична статистика, а саме — перевірка статистичних гіпотез за допомогою проведення статистичних експериментів.

Статистичний експеримент (або статистичне випробування) — це будь-яка процедура, яка може повторюватись безкінечну кількість разів і має чітко визначений набір можливих результатів, відомий як простір вибірки. Експеримент, який має більше одного можливого результату, називається випадковим. Кожен експеримент буде вважатись статистичним випробуванням, а результат цього випробування називатиметься випадковою подією.

Перевірка статистичних гіпотез — це процес перевірки статистичних припущень або гіпотез за допомогою аналізу даних, отриманих під час спостереження за статистичними експериментами.

Статистична гіпотеза — це будь-яке твердження, що описує певну сукупність. У процесі перевірки статистичних гіпотез формуються нульова (H0) та альтернативна (H1) гіпотези. Нульова гіпотеза має суперечити припущенню, яке ми хочемо вивчити (альтернативній гіпотезі). Важливо зазначити, що якщо наша альтернативна гіпотеза полягає в наявності певного зв’язку, то нульовою до неї буде “зв’язок не виявлено”, а не “зв’язок відсутній”. Тобто ми не підтверджуємо або спростовуємо гіпотезу, ми або відхиляємо її, або не можемо відхилитию. Така категоричність пов’язана з тим, що в будь-якому тестуванні лишається місце для помилки, і зв’язок, який ми не виявили, все ще може бути присутній, але, наприклад, бути слабшим, ніж ми припускали.

Статистичний розподіл — це спосіб опису розподілу значень змінної та ймовірностей, з якими ці значення трапляються. Статистичні розподіли використовуються для опису та вивчення великої кількості різних явищ і випадкових подій, а також для здійснення аналізу і прогнозування. Свій розподіл мають, зокрема, тривалість життя людини, зріст людини, похибки

вимірювальних апаратів і так далі. Кожен з існуючих розподілів має власні характеристики, що робить їх застосування корисним для різних типів аналізу даних.

Далі розглянемо тільки ті розподіли, які найчастіше траплятимуться саме в продуктовій аналітиці.

Основні параметри статистичного розподілу:

- Математичне сподівання (μ) — простими словами, середнє арифметичне значення розподілу;
- Стандартне відхилення (σ) та дисперсія — міри розсіювання даних навколо середнього значення. Дисперсія дорівнює σ^2;
- Мода — значення в розподілі, що зустрічається найчастіше;
- Медіана — значення, яке розташоване посередині розподілу і ділить його навпіл;
- Коефіцієнт асиметрії — міра, яка характеризує симетричність функції розподілу;
- Коефіцієнт ексцесу — міра, що характеризує “крутість” функції розподілу.

Крім цього, статистичні розподіли поділяються на дискретні (такі, в яких трапляються тільки цілі значення) та неперервні.

Нормальний розподіл (розподіл Ґауса) — це неперервний розподіл, який є одним із найбільш вживаних і важливих статистичних розподілів. Позначається $N(\mu, \sigma^2)$. Прикладами величин, що розподілені нормально, є зріст та вага людини, результати тестувань та екзаменів, температура повітря тощо.

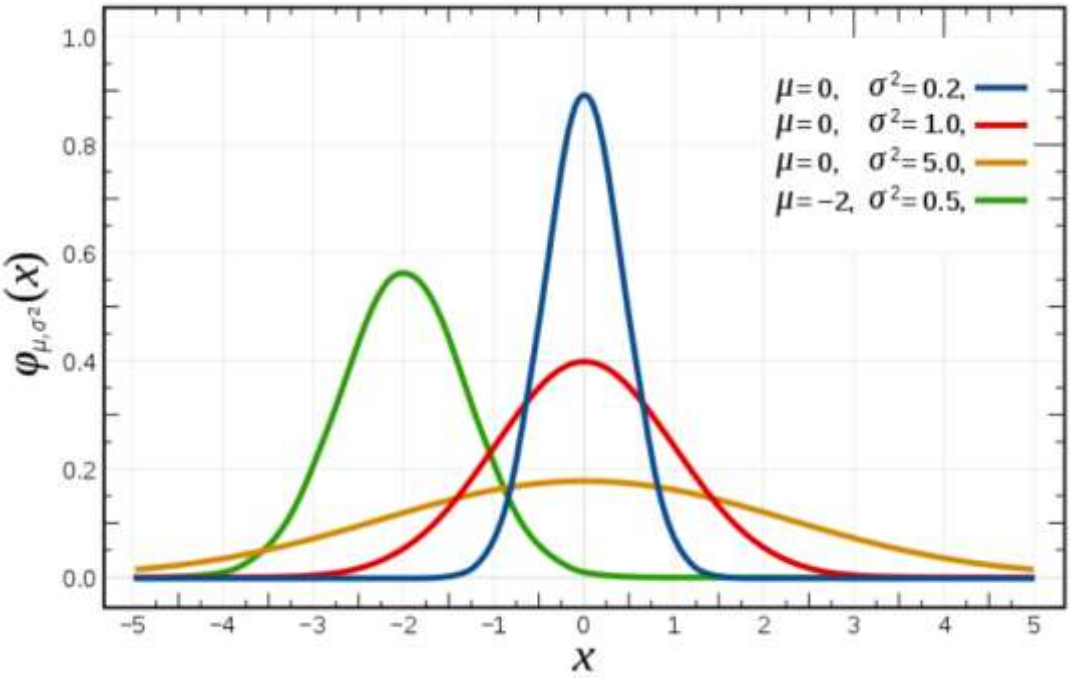


Рис. 1.8 Приклади розподілу Ґауса

Найважливіші характеристики:

1. Розподіл є одномодальним;
2. Математичне сподівання, мода та медіана розподілу — рівні:
3. Графік функції розподілу симетричний відносно математичного сподівання та має форму дзвону.

Окремим випадком нормального розподілу є стандартний нормальний розподіл — його математичне сподівання дорівнює нулю, а стандартне відхилення та дисперсія — одиниці.

Правило трьох σ — статистичне правило, що застосовується до нормального та наближених до нього розподілів і описує частку значень розподілу, що знаходяться в діапазонах ± σ, 2σ та 3σ відносно математичного очікування:

- В інтервал μ ± σ потрапляє 68,27% усіх значень;
- В інтервал μ ± 2σ — 95,45%;
- В інтервал μ ± 3σ — 99,73%.

Це правило часто використовується для роботи з викидами — значеннями, що суттєво відхиляються від загальної вибірки та трапляються з низькою ймовірністю.

Центральна гранична теорема (ЦРТ) полягає в тому, що коли ми беремо велику кількість незалежних та однорідно розподілених випадкових величин і обчислюємо їх середнє, то розподіл цих середніх буде наближатися до нормального розподілу навіть тоді, коли вихідні випадкові величини не мають нормального розподілу. ЦГТ дозволяє нам апроксимувати розподіли до нормального та аналізувати їх, користуючись властивостями нормального, як, наприклад, правило трьох σ.

Біноміальний розподіл — дискретний розподіл, який описує кількість успіхів у послідовності незалежних випробувань, результати яких можуть приймати лише два результати — успіх або невдача. Наприклад, біноміальний розподіл може описувати результати підкидання монетки, де випадання орла — це успіх, або конверсію в певну дію, тому біноміальний розподіл активно використовується в аналітиці. Іншою важливою рисою біноміального розподілу є те, що він є наближено нормальним.

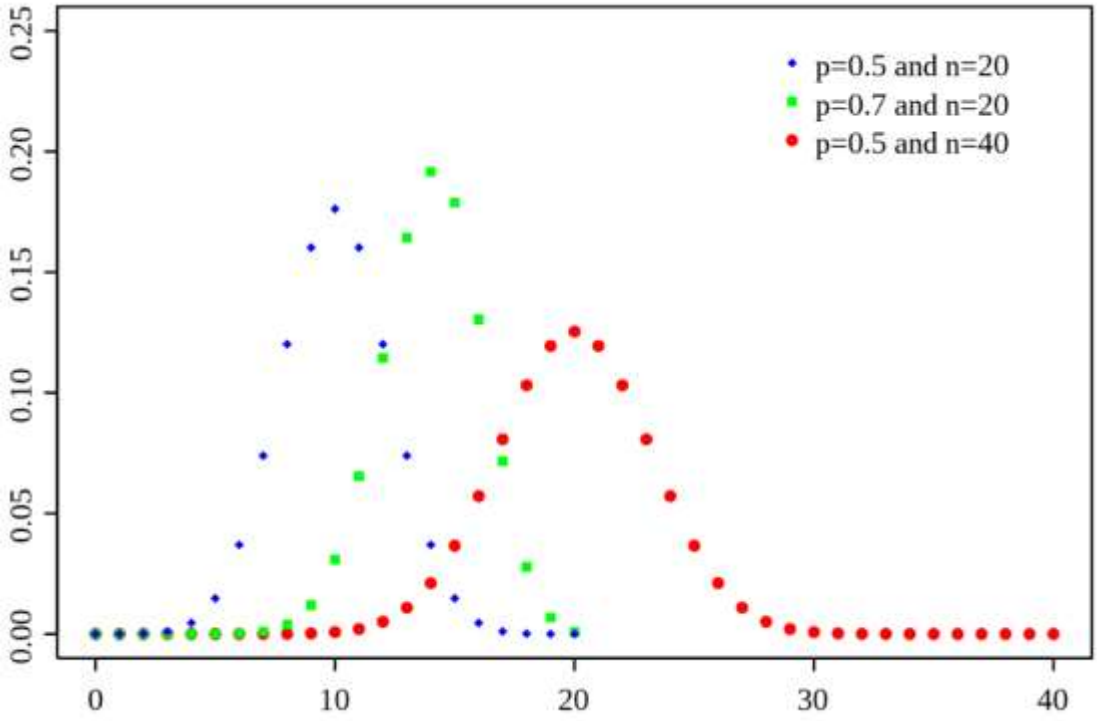


Рис. 1.9 Приклади біноміального розподілу

Біноміальний розподіл (Рис. 1.9) визначається кількістю випробувань (n) та ймовірністю успіху (p) та позначається B(n, p).

Найважливіші характеристики:

1. Математичне сподівання дорівнює np;

2. Якщо np — ціле число, тоді середнє, медіана і мода збігаються між собою і дорівнюють np;

3. Будь-яка медіана m обов'язково знаходиться в середині інтервалу $\lfloor np \rfloor \leq m \leq \lceil np \rceil$. Якщо p = 1/2 та n непарні, будь-яке число m в інтервалі $1/2(n - 1) \leq m \leq 1/2(n + 1)$ є медіаною біноміального розподілу. Якщо p = 1/2 і n парні, тоді m = n/2 є єдиною медіаною;

4. Мода розподілу дорівнює:

(n + 1)p, округлене до меншого цілого, якщо (n + 1)p = 0 або нецілe;

(n + 1)p та (n + 1)p - 1, якщо (n + 1)p є цілим числом і p не дорівнює нулю або одиниці;

n, якщо (n + 1)p = n + 1.

5. Якщо n є достатньо великим, тоді наближенням до B(n, p) буде N(np, np(1-p)).

Long-tailed розподіл, також відомий як розподіл Парето — розподіл, який характеризується наявністю довгого “хвоста”, який представляє менш поширені значення. Розподіл Парето визначає ймовірність того, що випадкова величина з розподілу є більшою за певне значення x та позначається P(X > x). Такий розподіл має населення міст, кількість підписників YouTube-каналів, перегляди тайтлів на Netflix тощо.

У розмовній версії цей розподіл частіше згадується як Закон Парето, який стверджує, що для багатьох явищ 80% наслідків спричинені 20% причин. Тобто довгий хвіст — це залишкові 20% наслідків, що спричинені 80% причин.

Long-tailed розподіл не є наближеним до нормального, тому для роботи з ним мають застосовуватись інші методи.

Закон великих чисел (ЗВЧ) — ще один важливий статистичний закон, що описує поведінку середніх значень великої кількості незалежних та однаково розподілених випадкових величин. Згідно з ЗВЧ, середнє значення вибірки має тенденцію збігатися з математичним сподіванням розподілу, до якого вона належить, коли розмір вибірки зростає до нескінченності.

Саме цей закон дозволяє нам припускати, що значення метрики випадкової вибірки, яка використовується для аналізу, відповідатиме значенню метрики для всієї популяції. Зокрема, це дозволяє нам приймати рішення на основі результатів A/B тесту.

Ми ознайомились з основними поняттями, що стосуються A/B тестування, і найбільш потрібними нам статистичними розподілами та законами, тепер про роботу з результатами A/B тестів.

Статистична значущість та p-value — це поняття, що використовуються для оцінки достовірності результатів досліджень. Статистична значущість результату в статистиці являє собою оцінку міри впевненості в його «істинності». У статистиці величину називають статистично значущою, якщо мала ймовірність чисто випадкового виникнення її або ще більш крайніх величин.

Рівень значущості тесту — це ймовірність ухвалити рішення відхилити нульову гіпотезу, якщо насправді вона вірна. Рівень значущості позначається буквою α та найчастіше дорівнює 10%, 5% або 1%.

p-value — це числове значення, яке описує ймовірність того, що отримані результати могли виникнути випадково, якщо нульова гіпотеза вірна. Чим нижче p-value, тим менша ймовірність хибно відхилити нульову гіпотезу. Результат вважається статистично значущим тоді, коли значення p-value менше за α.

У випадку A/B тестування, α фіксується на етапі підготовки тесту до запуску, а p-value визначається вже на етапі аналізу результатів і прийняття фінального рішення. Від виставленого рівня значущості напряму залежить

розмір вибірки, потрібний для проведення тесту: чим нижче значення α, тим менше ймовірність помилки і тим більший потрібний розмір вибірки.

Метод довірчих інтервалів — метод аналізу, який дозволяє оцінити значущість отриманого результату, базуючись на його належності до певного інтервалу, у якому, з певною ймовірністю, знаходиться реальне значення параметра. Це особливо корисно, коли ми працюємо з обмеженими або неповними даними та хочемо зробити висновки про популяцію загалом.

Довірчі інтервали розраховуються згідно з правилом трьох σ, з яким ми ознайомились раніше. Так, 95% довірчий інтервал розраховується як μ ± 1,96σ (приблизно два σ). Наприклад, якщо ми вимірюємо середній зріст людей із випадкової вибірки та розрахуємо до нього 95% довірчий інтервал, це означатиме, що середній зріст усіх людей з імовірністю 95% потраплятиме в цей інтервал.

Метод довірчих інтервалів можна використовувати для оцінки A/B тестів: якщо середнє значення вибірки A не належить до 95% довірчого інтервалу середнього значення вибірки B, то можемо вважати, що вибірка B відрізняється від вибірки A.

T-тест — статистичний критерій, що оцінює статистичну значущість різниці між середніми значеннями двох груп. Цей тест розроблений на основі розподілу Ст’юдента (t-розподілу) і є найпоширенішим методом порівняння середніх у статистичному аналізі. t-тест може застосовуватись тільки для даних, що розподілені нормально.

Існує кілька варіацій t-тесту:

- Independent Samples t-test (t-тест для незалежних вибірок) — використовується для порівняння середніх значень двох незалежних груп;
- Paired Samples t-test (t-тест для зіставлених вибірок) — використовується, коли в нас є зіставлені спостереження

(наприклад, перед вимірюванням і після) і ми хочемо перевірити, чи є статистично значуща різниця між цими спостереженнями;

- One-Sample t-test (t-тест для однієї вибірки) — використовується для визначення, чи є статистично значуща різниця між середнім значенням вибірки та певним значенням (наприклад, середнім значенням популяції).

Результатом t-тесу є t-статистика, яка співвідноситься зі значеннями з таблиці значень статистики, де число ступенів свободи розраховується як розмір вибірки - 1:

Критерій Ст'юдента *t*

Число ступенів свободи f	Рівень значимості α			
	0,10	0,05	0,01	0,001
1	6,31	12,70	63,70	637,00
2	2,92	4,30	9,92	31,60
3	2,35	3,18	5,84	12,90
4	2,13	2,78	4,60	8,61
5	2,01	2,57	4,03	6,86
6	1,94	2,45	3,71	5,96
7	1,89	2,36	3,50	5,40
8	1,86	2,31	3,36	5,04
9	1,83	2,26	3,25	4,78
10	1,81	2,23	3,17	4,59
11	1,80	2,20	3,11	4,44
12	1,78	2,18	3,05	4,32
13	1,77	2,16	3,01	4,22
14	1,76	2,14	2,98	4,14
15	1,75	2,13	2,95	4,07
16	1,75	2,12	2,92	4,01
17	1,74	2,11	2,90	3,96
18	1,73	2,10	2,88	3,92
19	1,73	2,09	2,86	3,88
20	1,73	2,09	2,85	3,85
21	1,72	2,08	2,83	3,82
22	1,72	2,07	2,82	3,79
23	1,71	2,07	2,81	3,77
24	1,71	2,06	2,80	3,74
25	1,71	2,06	2,79	3,72
26	1,71	2,06	2,78	3,71
27	1,71	2,05	2,77	3,69
28	1,70	2,05	2,76	3,66
29	1,70	2,05	2,76	3,66
30	1,70	2,04	2,75	3,65
40	1,68	2,02	2,70	3,55
60	1,67	2,00	2,66	3,46
120	1,66	1,98	2,62	3,37
∞	1,64	1,96	2,58	3,29

Рис 1.10 Критерій Ст'юдента

Тест Ст’юдента є популярним, проте не є єдиним чи універсальним критерієм оцінки.

ANOVA (Analysis of Variance, Дисперсійний аналіз) — це статистичний метод, який являє собою статистичний метод аналізу результатів, які залежать від якісних ознак. ANOVA порівнює дисперсії в межах трьох або більше груп з дисперсією між групами та дозволяє таким чином визначити, чи є між ними статистично значущі різниці.

ANOVA використовується для аналізу впливу факторів на залежну змінну або залежні змінні. Наприклад, якщо ми тестуємо три рекламні оголошення з різними слоганами, тоді фактором буде слоган на оголошенні, а залежною змінною — ефективність рекламного оголошення.

Результатом ANOVA є F-статистика (критерій Фішера), яка порівнює дисперсію між групами з дисперсією в межах груп. Якщо F-статистика є статистично значущою, це означає, що між групами існує статистично значуща різниця. Як і у випадку з критерієм Ст’юдента, F-статистика є табличним значенням:

Критерій Фішера *F*

Рівень значимості $\alpha = 0,01$								
$\backslash f_1$ / $f_2\backslash$	4	7	10	16	24	40	100	∞
1	5625,0	5928,0	6056,0	6169,0	6234,0	6286,0	6334,0	6366,0
2	99,25	99,34	99,40	99,44	99,46	99,48	99,49	99,50
3	28,71	27,67	27,23	26,83	26,60	26,41	26,23	26,12
4	15,98	14,98	14,54	14,15	13,93	13,74	13,57	13,46
5	11,39	10,45	10,05	9,68	9,47	9,29	9,13	9,02
6	9,15	8,26	7,87	7,52	7,31	7,14	6,99	6,88
7	7,85	7,00	6,62	6,27	6,07	5,90	5,75	5,65
8	7,01	6,19	5,82	5,48	5,28	5,11	4,96	4,86
9	6,42	5,62	5,26	4,92	4,73	4,56	4,41	4,31
10	5,99	5,21	4,85	4,52	4,33	4,17	4,01	3,91
12	5,41	4,65	4,30	3,98	3,78	3,61	3,46	3,36
14	5,03	4,28	3,94	3,62	3,43	3,26	3,11	3,00
16	4,77	4,03	3,69	3,37	3,18	3,01	2,86	2,75
18	4,58	3,85	3,51	3,19	3,00	2,83	2,68	2,57
Рівень значимості $\alpha = 0,05$								
$\backslash f_1$ / $f_2\backslash$	4	7	10	16	24	40	100	∞
1	225,0	237,0	242,0	246,0	249,0	251,0	253,0	254,0
2	19,25	19,36	19,39	19,43	19,45	19,47	19,49	19,50
3	9,12	8,88	8,78	8,69	8,64	8,60	8,56	8,53
4	6,39	6,09	5,96	5,84	5,77	5,71	5,66	5,63
5	5,19	4,88	4,74	4,60	4,53	4,46	4,40	4,36
6	4,53	4,21	4,06	3,92	3,84	3,77	3,71	3,67
7	4,12	3,79	3,63	3,49	3,41	3,34	3,28	3,23
8	3,84	3,50	3,34	3,20	3,12	3,05	2,98	2,93
9	3,63	3,29	3,13	2,98	2,90	2,82	2,76	2,71
10	3,48	3,14	2,97	2,82	2,74	2,67	2,59	2,54
12	3,26	2,92	2,76	2,60	2,50	2,42	2,35	2,30
14	3,11	2,77	2,60	2,44	2,35	2,27	2,19	2,13
16	3,01	2,66	2,49	2,33	2,24	2,16	2,07	2,01
18	2,93	2,58	2,41	2,25	2,15	2,07	1,98	1,92

Примітка: f_1 – відноситься до більшої дисперсії, f_2 – до меншої.

Рис. 1.11 Критерій Фішера

При використанні ANOVA ми робимо припущення, що розподіли є нормальними та незалежними, а також що різниця спостерігається лише між середніми, а дисперсія при цьому є однаковою.

Рівні ANOVA:

- One-Way ANOVA (однофакторний ANOVA) — використовується для порівняння середніх значень одного фактора або категорії між трьома або більше групами. Приклад: вплив різних вакцин від COVID-19 на захворюваність;
- Two-Way ANOVA (двофакторний ANOVA) — використовується для вивчення впливу двох факторів одночасно на залежну змінну. Приклад: вплив різних вакцин від COVID-19 на захворюваність окремо жінок та чоловіків. Факторами в цьому випадку є і вакцина, і стать пацієнта;
- Multivariate Analysis of Variance (MANOVA) — ANOVA, що дозволяє аналізувати вплив кількох незалежних змінних одночасно на кілька залежних змінних. Якщо продовжити розглядати приклад з вакциною, MANOVA дозволяє нам оцінити вплив різних вакцин на захворюваність і тяжкість перенесення хвороби окремо жінок та чоловіків. Тобто факторами знову є вакцина та стать, а залежними змінними — захворюваність і показник тяжкості перенесення захворювання.

Формування вибірки

Вибірка — це множина об'єктів або подій, вибраних за допомогою визначеної процедури з генеральної сукупності для участі в дослідженні. Коректне формування вибірки є однією з найважливіших складових проведення A/B тестів. Охопити всю генеральну сукупність під час A/B тестування в принципі не представляється можливим, тому ми використовуємо менші вибірки.

Випадкова вибірка (Random Sampling) — метод вибору об’єктів спостереження з популяції таким чином, щоб вони були “випадковими”,

тобто не відрізнялись від тих об’єктів, що до вибірки не потрапили. Існує кілька способів формування випадкової вибірки:

1. Проста (ймовірнісна) випадкова вибірка — згідно з цим методом, кожен об’єкт із генеральної сукупності може бути обраний з однаковою ймовірністю.

2. Стратифікована вибірка — цей метод полягає в розділенні генеральної сукупності на підгрупи за певною ознакою (страти) з подальшим випадковим обиранням вибірки з кожної групи у потрібній пропорції.

3. Кластерна вибірка — тут ми також ділимо генеральну сукупність на підгрупи, але в цьому випадку ми маємо дуже велику кількість підгруп і спочатку випадково обираємо кластери, які братимуть участь, а потім з них випадково обираємо об’єкти для фінальної вибірки;

4. Multistage (багатоетапна) вибірка — як стає зрозуміло з назви, це метод, згідно з яким вибірка формується в кілька етапів, кожен з яких може використовувати свій метод. Наприклад, спочатку ми формуємо групу за допомогою стратифікованої вибірки, а з неї вже формуємо просту випадкову вибірку.

Помилка виживання — коли обираються лише успішні суб’єкти, а неуспішні — ігноруються.

Помилки першого і другого роду — це поняття, тісно пов'язані з процесом статистичного тестування гіпотез. При проведенні A/B тестів важливо розуміти, що собою являють помилки першого та другого роду, з якою ймовірністю ми їх припускаємось та як це може повпливати на результат.

Помилка першого роду (Type I Error, α Error) — виникає, коли ми заявляємо про наявність статистично значущої різниці, коли її насправді немає. Помилку першого роду часто називають помилковою тривогою — наприклад, аналіз крові показав наявність захворювання в людини, хоча насправді людина не є хворою — спрацювала помилкова тривога.

Помилка другого роду (Type II Error, β Error) — виникає, коли ми не відхиляємо нульову гіпотезу та не реагуємо на ефект, хоча насправді він існує. Імовірність припуститись помилки другого роду визначає потужність β, проте, на відміну від рівня значущості, імовірність помилки дорівнюватиме не самій потужності, а 1 — β. Словом, чим вища потужність, тим менша ймовірність помилитись.

Важливо розуміти, що помилки першого і другого роду є неминучими в статистичних тестах і ми можемо лише зменшити їх імовірність. Зазвичай потужність і рівень значущості підбирають таким чином, щоб збалансувати ймовірності помилки, враховуючи особливості дослідження, яке ми проводимо: у деяких випадках помилка першого роду буде більш критичною для усунення, а в деяких — навпаки.

Окрім статистичних досліджень, помилки першого та другого роду також відіграють важливу роль у навчанні моделей-класифікаторів, які пишуть Data Science спеціалісти.

Порівнюючи метод аналізу ієрархій (MAI) в прийнятті рішень та спліт-тестування, можна сказати, що MAI більш затратний по ресурсам і складніший в автоматизації, тоді я к метод побудови і перевірки гіпотез через A/B тестування набагато зручніший в автоматизації і на його проведення витрачається суттєво менше часу. MAI доцільно використовувати тоді, коли є багато важелів впливу на прийняття рішення. В розрізі оцінки ефективності рекламних кампаній таких критеріїв зазвичай один або два – основна і додаткова метрика. Отже, приймаємо рішення на користь використання A/B тестування.

1.5 Постановка задачі на дипломний проєкт

Метою данного дипломного проєкту є розробка інформаційної панелі (дашборда) інструментами Google. Дані про роботу інтернет-магазину мають бути зібрані за допомогою GA4. В хмарному сховищі даних BigQuery потрібно очистити та відфільтрувати сирі дані. Після чого за допомогою Looker Studio потрібно розробити дашборд з основними метриками та воронкою продажів. Використовуючи A/B-тестування, зробити аналіз успішності двох або більше рекламних кампаній, проведених в рамках роботи інтернет-магазину.

РОЗДІЛ 2. Розробка моделі дашборду та схеми A/B тестування

2.1 Опис алгоритму отримання аналітичних метрик

Як відомо, корпорація Google – це компанія-флагман в індустрії інформаційних технологій. Безліч програмних продуктів були написані саме розробниками цієї компанії. Сьогодні ми щоденно користуємось Google-пошуком, Google-дисками, електронними таблицями Google Sheets і так далі.

Однак, є і такі програмні рішення, якими можна користуватись для проведення продуктової аналітики абсолютно безкоштовно. Ці інструменти достатньо потужні та надійні.

Для того, аби зацікавити молодих аналітиків, компанія Google створила можливість користуватися реальними даними з власного Google Market (https://shop.merch.google), який по суті є інтернет-магазином. Якби ми хотіли зібрати дані з власного інтернет-магазину, тоді ми би під'єднали наш сайт до GA4 і користувались би власними даними. Але в нашому випадку дані вже зібрані за допомогою GA4 і ми можемо їх вивчати в Google BigQuery.

Яку саме інформацію збирає GA4? Так як Google Market створено з метою продажу товарів з логотипом Google, то зрозуміло, що для подальшої аналітики потрібно знати, скільки користувачів потрапляє на сайт в кожний конкретний день, скільки з них ідуть без покупки, скільки кошиків так і не оплатили, скільки зробили покупку, який середній чек, чи повертався клієнт в інший день, яка вікова категорія зацікавилась товарами певної категорії, чи залежать продажі від пори року і так далі.

Розберемо структуру таблиці, в якій зберігаються дані. Кожен стовпець у таблиці events_РРРРММДД представляє параметр події. Стовпці таблиці описано нижче (Рис. 2.1 – Рис. 2.5).

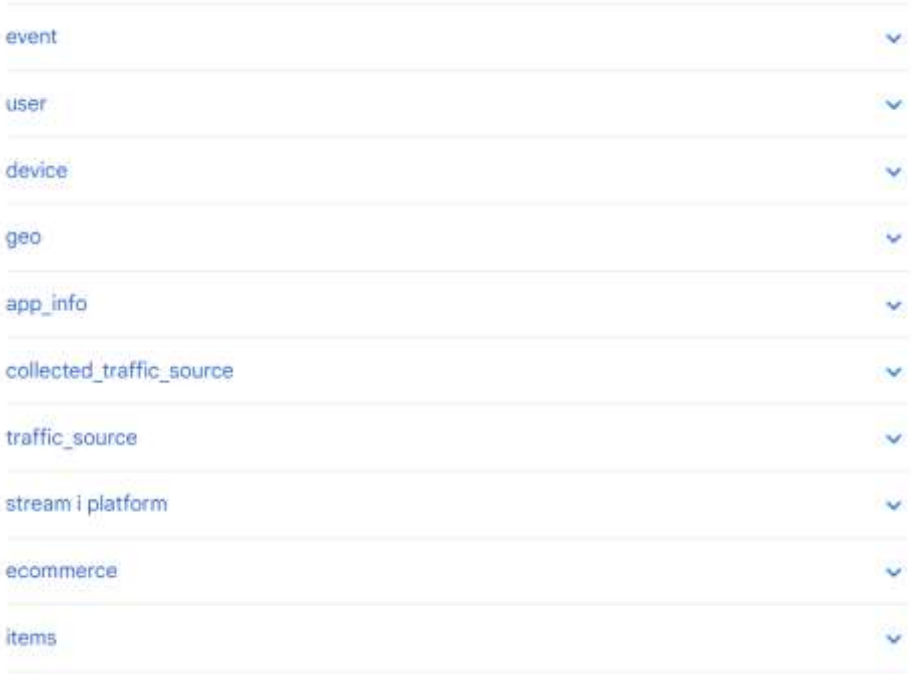


Рис. 2.1 Поля таблиці events_РРРРММДД

Поле event (подія) має наступні вкладені поля:

event

Ці поля містять інформацію, що є унікальною для події.

Назва поля	Тип даних	Опис
event_date	STRING	Дата реєстрації події в журналі (формат РРРРММДД в зареєстрованому часовому поясі вашого додатка).
event_timestamp	INTEGER	Час у мікросекундах (UTC), коли подію було зареєстровано в журналі клієнта.
event_previous_timestamp	INTEGER	Час у мікросекундах (UTC), коли подію було раніше зареєстровано в журналі клієнта.
event_name	STRING	Назва події.
event_value_in_usd	FLOAT	Значення параметра value події, виражене у валютному еквіваленті (у доларах США).
event_bundle_sequence_id	INTEGER	Послідовний ідентифікатор пакета, у якому завантажено ці події.
event_server_timestamp_offset	INTEGER	Зміщення позначки часу отримання й завантаження в мікросекундах.

Рис. 2.2 Вкладені поля для поля event (подія)

user

Ці поля містять інформацію, яка є унікальною для користувача, пов'язаного з подією.

Назва поля	Тип даних	Опис
is_active_user	ЛОГІЧНЕ ЗНАЧЕННЯ	Показник того, чи був користувач активним (True) або неактивним (False) у будь-який час протягом календарного дня. Включено лише в щоденні таблиці (`events_YYYYMMDD`).
user_id	STRING	Унікальний ідентифікатор, призначений користувачу.
user_pseudo_id	STRING	Псевдонімізований ідентифікатор (наприклад, ідентифікатор екземпляра додатка) користувача.
user_first_touch_timestamp	INTEGER	Час у мікросекундах, коли користувач уперше відкрив додаток чи відвідав сайт.

Рис. 2.3 Вкладені поля для поля user (користувач)

device

Цей ЗАПИС містить інформацію про пристрій, з якого відбулася подія.

Назва поля	Тип даних	Опис
device.category	STRING	Категорія пристрою (мобільний телефон, планшет, настільний комп'ютер).
device.mobile_brand_name	STRING	Назва бренду пристрою.
device.mobile_model_name	STRING	Назва моделі пристрою.
device.mobile_marketing_name	STRING	Маркетингова назва пристрою.
device.mobile_os_hardware_model	STRING	Інформація про модель пристрою, отримана безпосередньо з операційної системи.
device.operating_system	STRING	Операційна система пристрою.
device.operating_system_version	STRING	Версія ОС.
device.vendor_id	STRING	Ідентифікатор IDFV (лише якщо не отримано ідентифікатор IDFA).
device.advertising_id	STRING	Рекламний ідентифікатор або IDFA.
device.language	STRING	Мова ОС.
device.time_zone_offset_seconds	INTEGER	Відхилення часу в секундах відносно середнього часу за Гринвічем (GMT).

Рис. 2.4 Вкладені поля для поля device (пристрій)

geo

Цей запис містить інформацію про географічне місцеположення, де відбулася подія.		
Назва поля	**Тип даних**	**Опис**
geo.continent	STRING	Континент, на якому зареєстровано події (на основі IP-адреси).
geo.sub_continent	STRING	Субконтинент, на якому зареєстровано події (на основі IP-адреси).
geo.country	STRING	Країна, у якій зареєстровано події (на основі IP-адреси).
geo.region	STRING	Регіон, у якому зареєстровано події (на основі IP-адреси).
geo.metro	STRING	Муніципальний район, у якому зареєстровано події (на основі IP-адреси).
geo.city	STRING	Місто, у якому зареєстровано події (на основі IP-адреси).

Рис. 2.5 Вкладені поля для поля geo (локація користувача)

Дані про одну подію можуть відображатися в одному або кількох рядках залежно від того, чи містять вони повторювані записи. Наприклад, подія page_view з кількома параметрами event_params виглядатиме так, як указано в таблиці нижче. Перший рядок містить назву події, дату, позначку часу й інші об’єкти даних, які не повторюються (Рис. 2.6). Запис event_params повторюється для кожного параметра, зв’язаного з подією.

event_date	event_timestamp	event_name	event_params.key	event_params_value.strin
20220222	1643673600483790	page_view	page_location	https://example.com
			page_title	Головна сторінка
			medium	referral
			source	google
			page_referrer	https://www.google.com
			<параметри...>	*<значення...>*

Рис 2.6 Особливості зберігання записів в таблиці

Всі події (зафіксовані дії користувача сайту), які є в основній таблиці, можна отримати, використавши такий SQL-запит:

```
SELECT  distinct ds.event_name
FROM   `bigquery-public-data.ga4_obfuscated_ecommerce.events_*` AS ds
WHERE   _TABLE_SUFFIX BETWEEN ‘20240101’ AND ‘20240531’
```

Результат цього запиту буде наступним (Рис. 2.7):

Query results

JOB INFORMATION | RESULTS

Row	event_name
1	page_view
2	session_start
3	user_engagement
4	first_visit
5	scroll
6	view_item
7	view_search_results
8	add_payment_info
9	view_promotion
10	add_shipping_info
11	click
12	select_promotion
13	select_item
14	view_item_list
15	begin_checkout
16	purchase
17	add_to_cart

Рис. 2.7 Перелік дій користувача, які фіксуються на сайті.

Для подальшої аналітики будемо використовувати тільки наступні події: 'session_start', 'view_item', 'add_to_cart', 'begin_checkout', 'add_shipping_info', 'add_payment_info' та 'purchase'.

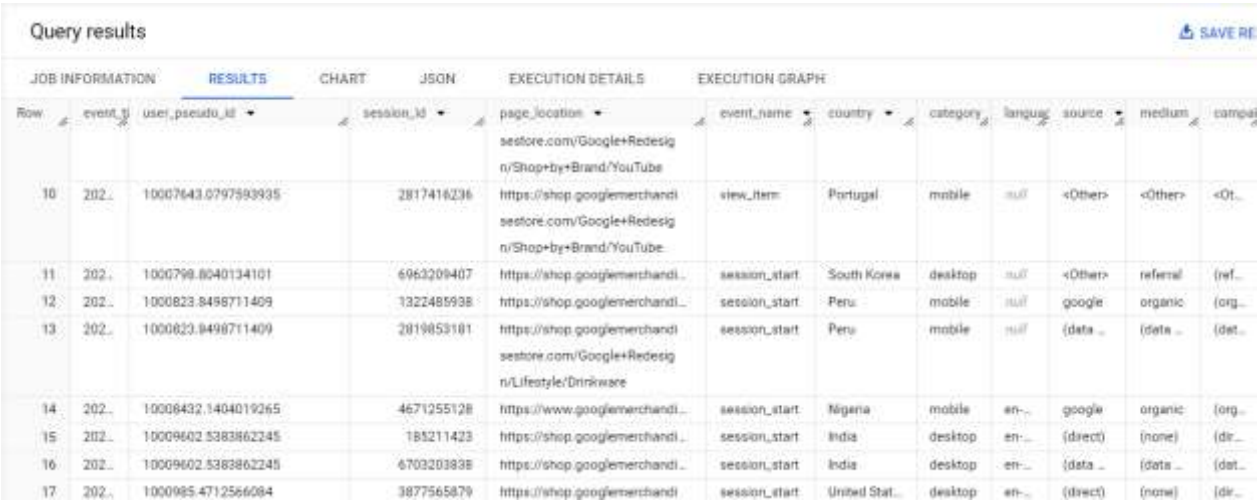
SQL-запит до таблиці GA4, в якій зберігаються записи про всі події (початок сесії користувача, огляд конкретного товару, додавання товару до кошика, початок формування замовлення, додавання інформації про доставку, внесення платіжної інформації і покупка) має такий вигляд:

```
with cte1 as
(SELECT ds.user_pseudo_id, --анонімний ідентифікатор користувача в GA4
```

```
(SELECT value.int_value FROM UNNEST(ds.event_params) WHERE key = 'ga_session_id') AS session_id, -- ідентифікатор сесії подій в GA4
(SELECT value.string_value FROM UNNEST(ds.event_params) WHERE key = 'page_location') AS page_location, --посадкова сторінка сесії
ds.geo.country, --країна користувача сайту
ds.device.category, --категорія пристрою користувача
ds.device.language, --мова пристрою користувача
ds.traffic_source.source, -- джерело відвідування сайту
ds.traffic_source.medium, --medium відвідування сайту
ds.traffic_source.name as campaign, --назва кампанії відвідування сайту
concat(ds.user_pseudo_id, '/', (SELECT value.int_value FROM UNNEST(ds.event_params) WHERE key = 'ga_session_id')) as userId_session
FROM `bigquery-public-data.ga4_obfuscated_sample_ecommerce.events_*` as ds
where ds.user_pseudo_id is not null
and ds.event_name = 'session_start'),
cte2 as
(SELECT timestamp_micros(ds.event_timestamp) as event_timestamp , --дата та час події
ds.event_name, --назва події
concat(ds.user_pseudo_id, '/', (SELECT value.int_value FROM UNNEST(ds.event_params) WHERE key = 'ga_session_id')) as userId_session
FROM `bigquery-public-data.ga4_obfuscated_sample_ecommerce.events_*` as ds
where ds.user_pseudo_id is not null
and ds.event_name in ('session_start', 'view_item', 'add_to_cart', 'begin_checkout', 'add_shipping_info','add_payment_info', 'purchase'))
select cte2.event_timestamp, cte1.session_id, cte1.user_pseudo_id, cte1.page_location, cte2.event_name, cte1.country,
cte1.category, cte1.language, cte1.source,  cte1.medium, cte1.campaign, cte1.userId_session
from cte1
left join cte2
on cte1.userId_session=cte2.userId_session
```

SQL-запит доволі складний і має дві тимчасові вибірки cte1 і cte2, які потім об’єднуються через left join і зв’язуються за полем, яке вміщає ідентифікатор користувача і ідентифікатор його поточної сесії на сайті. Ці особливості запиту обумовлені тим, що дані про події на сайті не завжди фіксуються коректно, тому нам потрібно виключити ті записи, де нема старту сесії, а є покупка, наприклад. Тобто на цьому етапі ми очищуємо дані від помилкових і неповних записів.

Фрагмент даних буде мати такий вигляд:



Query results

SAVE RE

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	event_ti	user_pseudo_id	session_id	page_location	event_name	country	category	languag	source	medium	campai
				sestore.com/Google+Redesig n/Shop+by+Brand/YouTube							
10	202..	10007643.0797593935	2817416236	https://shop.googlemerchandi sestore.com/Google+Redesig n/Shop+by+Brand/YouTube	view_item	Portugal	mobile	null	<Other>	<Other>	<Ot..
11	202..	1000798.8040134101	6963209407	https://shop.googlemerchandi..	session_start	South Korea	desktop	null	<Other>	referral	(ref..
12	202..	1000823.8498711409	1322485938	https://shop.googlemerchandi..	session_start	Peru	mobile	null	google	organic	(org..
13	202..	1000823.8498711409	2819853101	https://shop.googlemerchandi sestore.com/Google+Redesig n/Lifestyle/Drinkware	session_start	Peru	mobile	null	(data ..	(data ..	(dat..
14	202..	10008432.1404019265	4671255128	https://www.googlemerchandi..	session_start	Nigeria	mobile	en-..	google	organic	(org..
15	202..	10009602.5383862245	185211423	https://shop.googlemerchandi..	session_start	India	desktop	en-..	(direct)	(none)	(dir..
16	202..	10009602.5383862245	6703203838	https://shop.googlemerchandi..	session_start	India	desktop	en-..	(data ..	(data ..	(dat..
17	202..	1000985.4712566084	3877565879	https://shop.googlemerchandi	session_start	United Stat..	desktop	en-..	(direct)	(none)	(dir..

Рис 2.8 Дані для дашборду

Тепер, маючи очищені та підготовлені дані, можна будувати дашборд з метриками, діаграмами та комерційною воронкою.

2.2 Опис інтерфейсу користувача

Для побудови інформаційної панелі (дашборду) використовувався такий аналітичний Google- інструмент як Looker Studio. Це середовище дає змогу перетворювати дані на зрозумілі інформативні звіти, якими легко обмінюватися та які можна персоналізувати. Ось декілька основних відмінностей Looker Studio, які можуть бути перевагами перед статичними звітами:

- можливість легко підключити різні джерела даних (як на локальному пристрої, так і в хмарному сховищі);
- можливість візуалізувати дані за допомогою інтерактивних діаграм і таблиць;
- можливість поділитися статистикою зі своєю командою або з ким завгодно.
- можливість працювати разом декільком людям над звітами.
- можливість пришвидшити процес створення звітів, використовуючи вбудовані кольорові теми.

Аналітична інформаційна панель «Комерційна воронкаshop.merch.google» (Рис. 2.9, Рис. 2.10) складається з таких метрик:

1. показники кількості користувачів, сесій, замовлень та покупок;
2. таблиця джерел потрапляння на сайт;
3. таблиця рівня конверсії в розрізі мов пристроїв;
4. таблиця рейтингу стартових сторінок;
5. теплова мапа ко кількості відкритих сесій в різних країнах світу;
6. кругова діаграма в розрізі пристроїв користувачів;
7. комерційна воронка продажів.

Рис. 2.9 Інформаційна панель. Частина 1

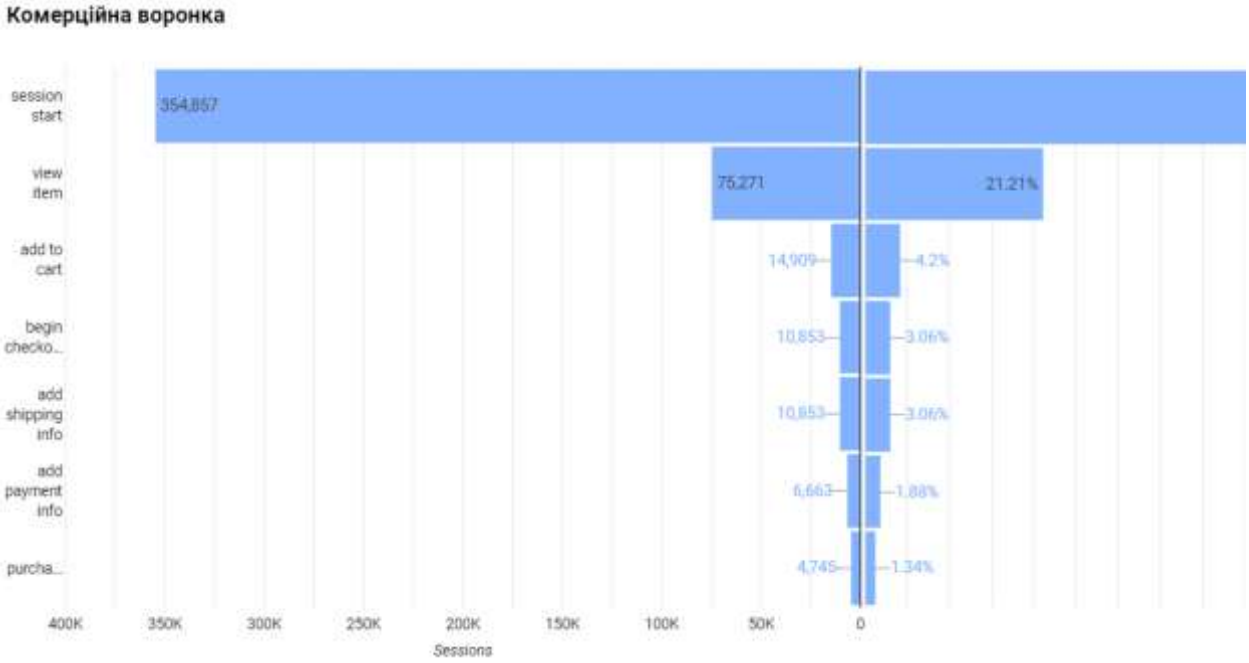


Рис. 2.10 Інформаційна панель. Частина 2

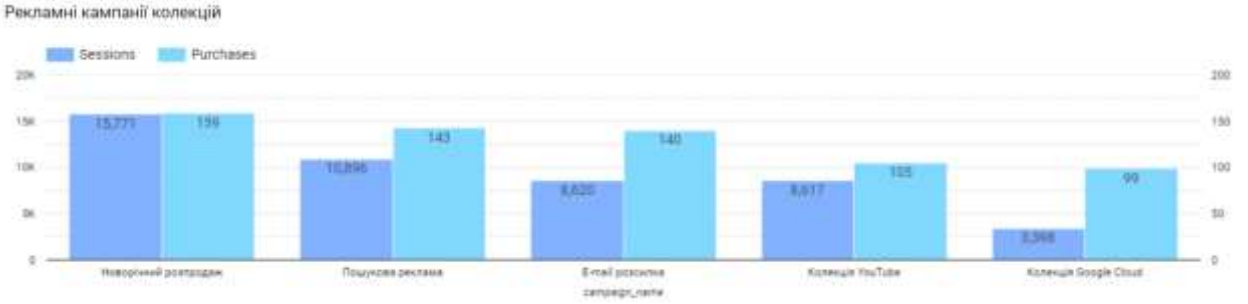


Рис. 2.11 Інформаційна панель. Частина 3

Підключення джерела даних в Looker Studio має наступний вигляд (Рис 2.12) і тут прописується sql-запит в BigQuery (Рис 2.13):

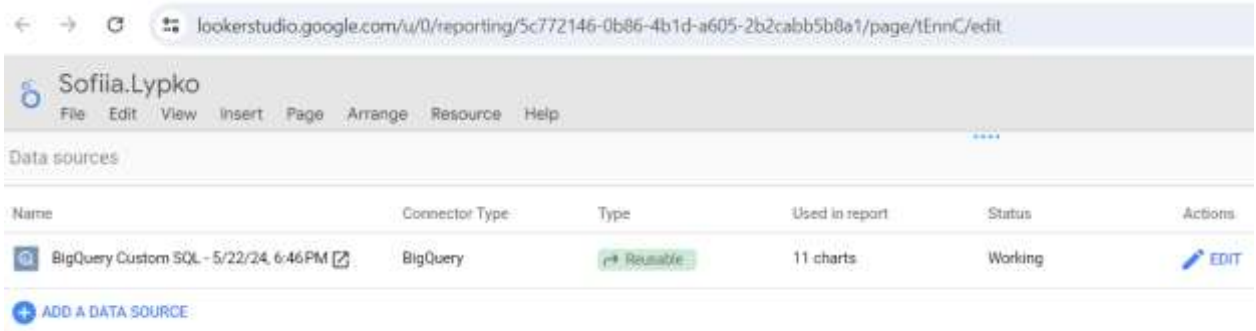


Рис. 2.12 Додавання набору даних з BigQuery

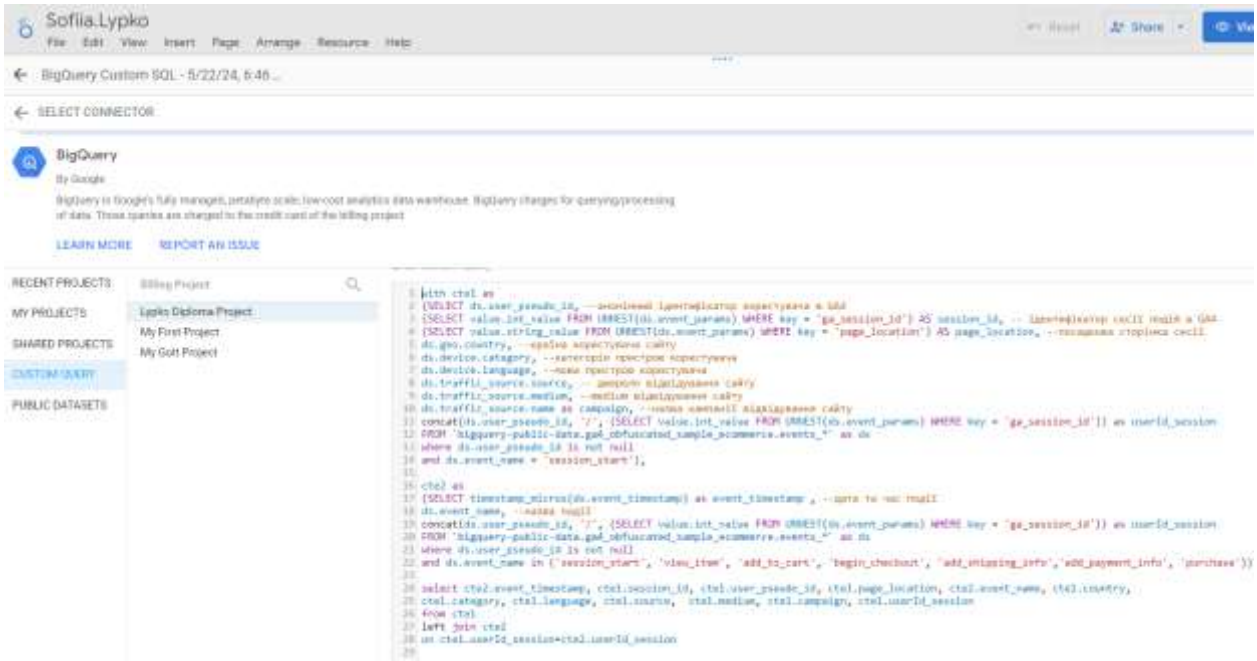


Рис. 2.13 SQL-запит з BigQuery

Після підключення набору даних до дашборду в Looker Studio можна додавати різні елемени на панель і прописувати їх обчислення. Таким чином, показники кількості користувачів, сесій, замовлень та покупок обчислюються як кількість унікальних значень по полям user_pseudo_id , session_id, для підрахунку кількості замовлень прописується фітр (Рис. 2.14), де рахуємо тільки події add to cart (додавання товару в кошик), а для кількості покупок

прописується фільтр (Рис. 2.15), де враховуються події лише purchase (покупка).

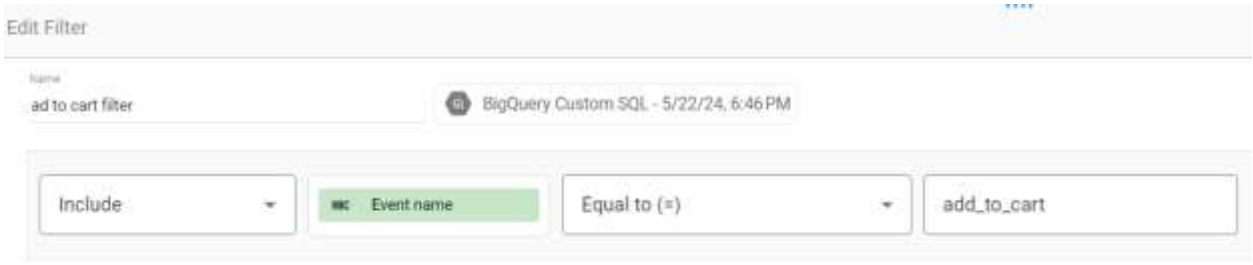


Рис. 2.14 Фільтр події add to cart (додавання товару в кошик

Рис. 2.15 Фільтр події purchase (покупка)

Таблиця «Джерела потрапляння на сайт» (Рис. 2.9) має стовпчики (за якими групуються дані) Sourse, Medium, Campaign і стовпчик сумарної кількості унікальних сесій Sessions (Рис. 2.16).

Рис. 2.16 Налаштування стовпчиків таблиці «Джерела потрапляння на сайт»

Поле Source містить назву рекламного майданчика, на якому було розміщене посилання, наприклад, Twitter, Google Ads, Facebook, Instagram або блог.

Поле Medium допомагає визначити тип каналу, з якого приходить трафік: покупці перейшли на сайт із контекстної реклами з оплатою за клік, банерної реклами з оплатою за покази або розсилки.

Список найбільш популярних значень для поля medium:

- cpc (cost per click) або ppc (pay per click) — контекстна реклама з оплатою за клік;
- social_cpc — реклама в соціальних мережах з оплатою за клік;
- display — банерна реклама з оплатою за покази;

- referral — перехід з іншого вебресурсу;
- email — розсилка;
- organic — безкоштовний пошук.

Поле Campaign використовується для назви кампанії, продукту або ключового слова. Цей параметр допомагає ідентифікувати різні рекламні кампанії у статистиці. Приклади значень: valentine, 8_march, mothers_day.

Таблиця «Рівень конверсії за мовою пристрою» (Рис. 2.9) відображає долю користувачів, які здійснили покупку, в розрізі мов пристроїв користувачів. Оскільки в базі зберігаються скорочені назви мов, довелось додати обчислюване поле, щоб візуально адаптувати дані для сприйняття. Код обробки поля Language на мові SQLмає такий вигляд:

```
case t0._language_
when 'en-us' then 'English'
when 'en-gb' then 'English'
when 'en-ca' then 'English'
when 'en' then 'English'
when 'zh' then 'Chinese'
when 'fr' then 'French'
when 'de' then 'German'
when 'ko' then 'Korean'
when 'es-se' then 'Spanish'
else  'Not defined'
end
```

В таблиці також присутнє поле кількості сесій, кількості покупок та CR (рівень конверсії), який обчислюється діленням кількості покупок на кількість сесій. Конверсія – важливий показник ефективності маркетингових кампаній. Що вона вища, то більше відвідувачів сайту виконують цільові дії. Це означає, що маркетингові кампанії працюють продуктивно та приводять до збільшення продажів і оборотів бізнесу.

На конверсію можуть впливати різні чинники, наприклад новизна товарів, ціна, інтерфейс сайту, дратуючий чи заспокійливий контент, акції, а також якість стратегії маркетингу та взаємодії з цільовою аудиторією. Її вимірювання й аналіз допомагають визначити ефективність стратегії продажів, щоб внести необхідні зміни для покращення торгівлі.

Рівень конверсії зазвичай коливається від 1 до 5%. Однак деякі інтернет-магазини можуть досягати вищих показників, якщо пропонують високоякісний продукт, конкурентну ціну, зручний сайт і ефективну маркетингову стратегію. Наприклад, для інтернет-магазинів, котрі спеціалізуються на продажах унікальних або ексклюзивних товарів, конверсія може перевищувати 10%. А в інтернет-магазинах, які продають продукцію масового попиту, цей показник часто падає нижче 1%.

Наступна діаграма на дашборді (Рис. 2.9) – теплова мапа. По будь-якій країні світу можна побачити кількість користувачів. Найбільш активні країни позначені темними відтінками. Так як дашборд інтерактивний, то користувач може клікнути по будь-якій країні на мапі і всі діагарами й таблиці перерахуються в розрізі саме цієї країни.

Наступна таблиця – «Рейтинг стартових сторінок». В ній дані згруповані по сайтам, з яких користувачі прийшли в інтернет-магазин. Дані впорядковані за кількістю унікальних сесій.

Кругова діаграма на дашборді (Рис. 2.9) відображає кількість сесій з розбивкою по пристроям користувачів: телефону, планшету або комп’ютеру.

Ключова діаграма даного дашборду – «Комерційна воронка» (Рис. 2.10), яка складається з двох стовпчастих діагарам. Обидві частини ілюструють послідовність дій користувачів від потрапляння на сайт магазину до здійснення покупки і сформовані в дзеркальному відображенні, щоб відтворити саме вигляд воронки. Ліва частина показує кількість користувачів, які виконали бажану дію на кожному етапі воронки

(наприклад, додали товар в кошик) в абсолютних числах, а права частина – у відсотках.

Стовбчаста діагарама, яка відображає кількість сесій та кількість покупок в розрізі маркетингової кампанії (Рис. 2.12) будемо використовувати як приклад A/B-тестування для прогнозування успішності цих кампаній в подальшому.

2.3 Розрахунок ефективності кампаній за методом аналізу ієрархій

Розглянемо метод аналізу ієрархій, який дозволяє кількісно визначити порівняльну важливість критеріїв та субкритеріїв оцінки маркетингової ефективності проведених кампаній. Цей метод припускає проведення попарних порівнянь об’єктів з використанням суб'єктивних суджень, чисельно оцінюваних за визначеною шкалою.

Критерії з найбільшими величинами важливості доцільно використовувати при розробці подальшої стратегії конкурентної політики інтернет-магазину. Перевагою даного методу є визначення не тільки порядку пріоритетів кожного окремого критерію, але і величини пріоритету.

1 етап. Декомпозиція проблеми.

Ми маємо дві рекламні кампанії в інтернет-магазині: Колекції YouTube і Колекції Google Cloud. Фокус нашої проблеми полягає у визначенні найбільш ефективної кампанії. Для кожного виду кампаній приймаємо наступні критерії: охоплення, конверсія і дохід.

2 етап. Представлення проблеми у вигляді ієрархії (Рис. 2.17).

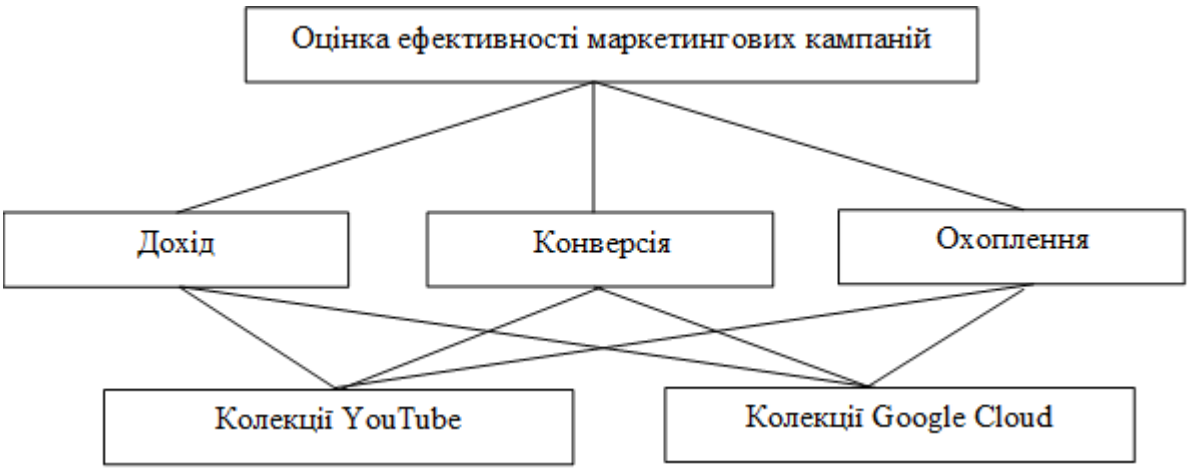


Рис. 2.17 Ієрархічне представлення компонентів проблеми

3 етап. Експертне оцінювання переваг.

На цьому етапі ми встановлюємо пріоритети кожного з трьох критеріїв по відношенню до інших (Таблиця 2.1). Зрозуміло, що відношення елементу самого до себе дорівнює 1. Елементи матриці визначаться за шкалою Сааті.

Таблиця 2.1 Матриця попарних порівнянь критеріїв

	Охоплення	Конверсія	Дохід
Охоплення	1	3	5
Конверсія	1/3	1	2
Дохід	1/5	1/2	1

4 етап. На 4 етапі обчислюємо вектор пріоритетів за матрицею попарних порівнянь. Приблизні оцінки значення головного власного вектора можна отримати чотирма способами, ми обираємо “Середньогеометричні значення за рядком”. Для цього необхідно перемножити n елементів кожного рядка і витягти корінь n-го ступеня з отриманого (Таблиця 2.2).

Таблиця 2.2 Вектор пріоритетів

	Охоплення	Конверсія	Дохід	П	$\sqrt[3]{П}$	w
Охоплення	1	3	5	15	2.47	**0.65**
Конверсія	1/3	1	2	0.67	0.88	**0.23**
Дохід	1/5	1/2	1	0.1	0.46	**0.12**
Σ					3.81	**1**

Спочатку формуємо таблицю показників кампаній за кожним з критеріїв (в нашому випадку дохід від продажу кожної з колекцій – це умовні суми, бо реальні доходи комерційні компанії в такому розрізі не надають на загал).

Таблиця 2.3 Показники кампаній за критеріями

	Охоплення	Конверсія	Дохід
Колекції YouTube (YTC)	8617	0.01	20000
Колекції Google Cloud (GCC)	3398	0.03	28000

Далі визначаємо локальні пріоритети по кожному з критеріїв та їх вектор (Табл. 2.4 – Табл. 2.6). Де П – добуток пріоритетів по різних кампаниях, $\sqrt[3]{П}$ корень третьої степені з добутку П, значення w визначається як доля в загальній сумі.

Таблиця 2.4 Матриця попарних порівнянь за критерієм «Охоплення»

Охоплення	GCC	YTC	П	$\sqrt[3]{П}$	w
GCC	1	1/3	0.33	0.69	0.32
YTC	3	1	3	1.44	0.68
Σ				2.13	1

Таблиця 2.5 Матриця попарних порівнянь за критерієм «Конверсія»

Конверсія	YTC	GCC	П	$\sqrt[3]{П}$	p(2)
YTC	1	1/4	0.25	0.63	0.28
GCC	4	1	4	1.59	0.72
Σ				2.22	1

Таблиця 2.6 Матриця попарних порівнянь за критерієм «Дохід»

Дохід	GCC	YTC	П	$\sqrt[3]{П}$	p(3)
GCC	1	1/2	0.5	0.79	0.39
YTC	2	1	2	1.26	0.61
Σ				2.05	1

На наступному етапі застосовуємо принцип синтезу для визначення глобальних пріоритетів. Глобальні пріоритети елементів визначаються, як сума додатків локальних пріоритетів на глобальні пріоритети елементів (Таблиця 2.7). Тобто, оцінка GCC=0,65*0,32+0,23*0,72+0,12*0,39.

Таблиця 2.7 Оцінка результату

	Охоплення	Конверсія	Дохід	
w	0.65	0.23	0.12	Оцінка
GCC	0.32	0.72	0.39	0.42
YTC	0.68	0.28	0.61	**0.58**

Етап інтерпретації та аналіз результатів. З огляду на пріоритетність критеріїв оцінки, проводимо оцінку маркетингових кампаній. Результати дослідження свідчать про те, що більш високий рівень ефективності спостерігається у маркетингової кампанії Колекції YouTube (YTC) – 0,58 в порівнянні з кампанією Колекції Google Cloud (GCC).

2.4 Підготовка та проведення A/B тестуівння

Для проведення спліт-тестування потрібні дві групи користувачів, тож в нас буде група А – користувачі, яким показали рекламу Колекції YouTube і група В – корстувачі, що побачили рекламу Колекції Google Cloud (Рис 2.11).

Статистичну обробку даних будемо проводити на мові програмування Python, це найзручніший інструмент на сьогоднішній день.

Для метематичних обчислень необхідно доєднати модуль SciPy. SciPy (Scientific Python) — це бібліотека для Python, яка надає набір інструментів для виконання наукових та інженерних обчислень. SciPy побудована на основі бібліотеки NumPy та складається з кількох модулів, із яких нам буде потрібний модуль для статистичних обчислень — scipy.stats.

scipy.stats надає широкий спектр функцій для статистичного аналізу, наприклад:

- функції для роботи за статистичними розподілами;
- функції описової статистики;
- функції для тестування гіпотез тощо.

За допомогою функцій тестування гіпотез із модуля scipy.stats ми маємо змогу оцінити статистичну значущість результатів нашого A/B тесту.

Вхідні дані:

Контрольна група А

Кількість користувачів у групі: 8617

Кількість конверсій у групі: 105

Значення конверсії: 1.22%

Альтернативна група В

Кількість користувачів у групі: 3398

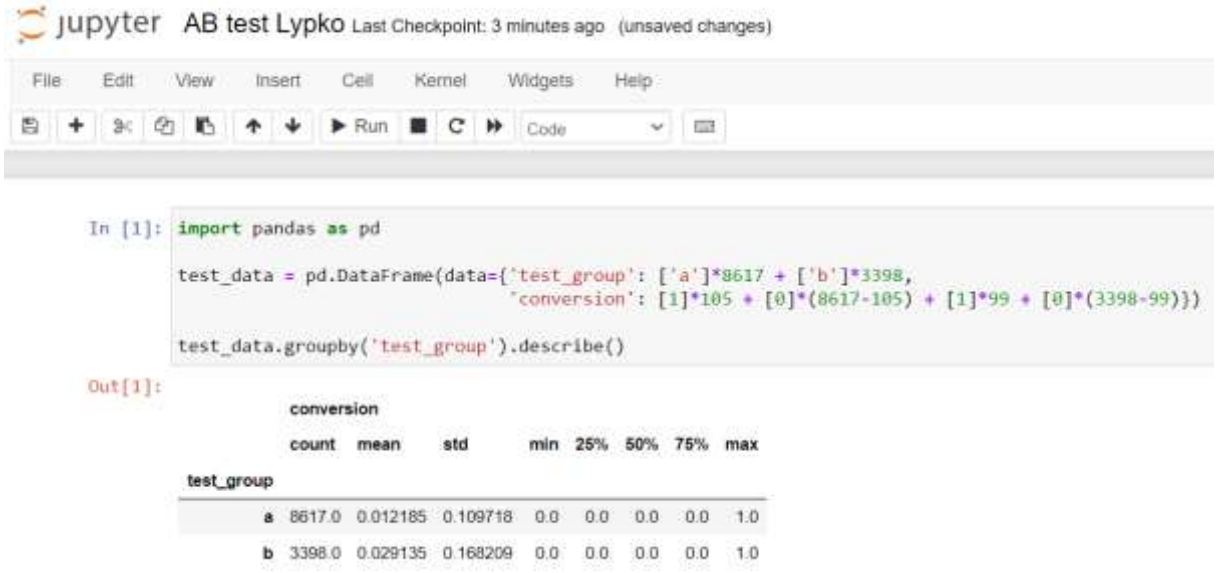
Кількість конверсій у групі: 99

Значення конверсії: 2.91%

Для простоти відтворення ми згенеруємо датасет для подальшої роботи самостійно, для цього нам потрібно створити DataFrame з двома колонками, в якому кожен рядок позначає одного користувача:

test_group має містити дані про те, до якої групи належить користувач — a або b;

conversion — містить дані про наявність конверсії в користувача, 1 — якщо користувач здійснив конверсію і 0 — якщо ні.



	conversion							
	count	mean	std	min	25%	50%	75%	max
test_group								
a	8617.0	0.012185	0.109718	0.0	0.0	0.0	0.0	1.0
b	3398.0	0.029135	0.168209	0.0	0.0	0.0	0.0	1.0

Рис. 2.18 Результат виконання коду

Значення mean у таблиці (Рис. 2.18) — це значення конверсії в наших групах, а отже, все згенеровано коректно.

Для тестування гіпотез за допомогою SciPy будемо використовувати Критерій Ст’юдента:

scipy.stats.ttest_ind — Тест (Тест Ст’юдента) для двох незалежних вибірок

Нульова гіпотеза: середні величини двох незалежних вибірок не відрізняються.

Основні параметри функції:

a, b — масиви значень у двох вибірках;

axis — напрямок обчислень у випадку використання багатовимірних масивів;

equal_var — параметр, який вказує, чи припускається рівність дисперсій. За замовчуванням, true;

alternative — визначає альтернативну гіпотезу.

less — середнє вибірки A менше за середнє вибірки B;

greater — середнє вибірки A більше за середнє вибірки B;

two-sided — середні вибірок відрізняються.

У нашому випадку ми вже знаємо, що конверсія в групі B вища, а також завдяки функції describe знаємо, що умова рівності дисперсій не виконується, оскільки стандартні відхилення в групах різні, тож чи є це покращення статистично значущим, якщо ми задамо α = 0.05?

```
In [2]: from scipy import stats

        alpha = 0.05

        statistic, pvalue = stats.ttest_ind(test_data[test_data['test_group'] == 'a']['conversion'],
                                            test_data[test_data['test_group'] == 'b']['conversion'],
                                            alternative='less')

        print(f't-statistic: {round(statistic, 2)}, p-value: {round(pvalue, 2)}')

        if pvalue < alpha:
            print('Різниця статистично значуща, нульова гіпотеза відхиляється.')
        else:
            print('Різниця не є статистично значущою, нульову гіпотезу не можна відхилити.')

        t-statistic: -6.49, p-value: 0.0
        Різниця статистично значуща, нульова гіпотеза відхиляється.
```

Рис. 2.19 Аналіз різниці за допомогою аналітичних модулів Python

Отже, за результатом виконання коду бачимо, що Різниця тестування є статистично значущою, тому нульову гіпотезу відхиляємо. Можемо приймати альтернативний варіант, оскільки він статистично значуще кращий за контрольний, тобто рекламна кампанія «Колекція Google Cloud» є значно успішнішою за кампанію «Колекція YouTube».

Розглянемо інший варіант тестування - тести з перестановками.

Permutation test, або тест з перестановками — метод статистичного тестування, який базується на перемішуванні даних між групами для оцінки статистичної значущості результатів.

Основна ідея полягає в тому, щоб за допомогою перестановок зрозуміти, з якою ймовірністю ми отримаємо такі самі або більш екстремальні результати, якщо нульова гіпотеза справджується. Для цього проводиться багато перемішувань, для кожного з яких обчислюється статистика, а потім ми порівнюємо фактичну статистику з отриманим розподілом.

Алгоритм, якщо обирати критерій Ст’юдента, виглядає приблизно так:

Розраховуємо t-статистику на тих даних, які маємо. Це і буде наша фактична статистика t0;

Тепер об’єднаємо всі спостереження в одну групу й розділимо на дві рівних заново випадковим чином;

Розрахуємо t-статистику на цих нових двох групах;

Повертаємось до кроку 2, потім — до кроку 3. Робимо все те ж саме n разів;

Тепер усі отримані t-статистики упорядковуємо у зростаючому порядку — це наш емпіричний розподіл.

Тепер, якщо наше t0 не потрапляє в середні 95% значень емпіричного розподілу, ми можемо відкидати нульову гіпотезу з імовірністю 95%.

SciPy пропонує нам використовувати для тестів з перестановками вже готову функціюscipy.stats.permutation_test.

Нульова гіпотеза: дані рандомно обрані з одного розподілу.

Основні параметри функції:

data — дані, що будуть використовуватися для перестановок;

statistic — функція, що повертатиме статистику, яку ми хочемо використовувати для перестановок. Наприклад, t-статистику для критерію Ст’юдента;

n_resamples — кількість перемішувань;

alternative — визначає альтернативну гіпотезу, аналогічно до того, як це працює в інших тестах.

Наприклад, спробуємо тест із сотнею перестановок та використанням критерію Ст'юдента на тих же даних, що й в попередньому блоці:

```
In [3]: from scipy import stats

        def statistic(x, y):
            return stats.ttest_ind(x, y).statistic

        alpha = 0.05

        x = test_data[test_data['test_group'] == 'a']['conversion']
        y = test_data[test_data['test_group'] == 'b']['conversion']

        results = stats.permutation_test((x, y), statistic, n_resamples=100)

        print(f'statistic: {round(results.statistic, 2)}, p-value: {round(results.pvalue, 2)}')

        if results.pvalue < alpha:
            print('Різниця статистично значуща, нульова гіпотеза відхиляється.')
        else:
            print('Різниця не є статистично значущою, нульову гіпотезу не можна відхилити.')

        statistic: -6.49, p-value: 0.02
        Різниця статистично значуща, нульова гіпотеза відхиляється.
```

Рис. 2.20 Тест з перестановками

По результатам тесту з перестановками бачимо, що слід відхилити нульову гіпотезу та робимо висновок, що різниця між нашими групами A і B є статистично значущою.

2.5 Оцінка результатів і побудова гіпотез підвищення ефективності кампаній

Висновки, які можна зробити по даному дашборду, очевидні:

1. Найбільша кількість користувачів потрапляє на сайт магазину через безкоштовний пошук товарів в Google або напряму в інтернет-магазин.
2. Абсолютна більшість користувачів мають англомовні пристрої.
3. Найбільш зацікавлені в покупках користувачі проживають в США та Канаді.
4. Майже ніхто не здійснює покупок на планшетах.
5. Приблизно 80% користувачів «відпадають» навіть до огляду конкретного товару.
6. Конверсія відвідувачів сайту в покупців складає 1.34%, що вважається непоганим результатом.

Працювати над їдеями покращеня конверсії або над ідеями залучення нових користувачів будуть менеджери разом з маркетологами. Але дашборд може підказати напрямки, за якими слід рухатись.

Щодо результатів A/B тестування, то їх варто візуалізувати.

```
In [5]: import matplotlib.pyplot as plt
        import seaborn as sns

        plt.figure(figsize=(8, 6))
        sns.barplot(x=test_data['test_group'],
                    y=test_data['conversion'],
                    errorbar=('se'))

        plt.title('Mean Comparison with Standart Error')
        plt.xlabel('Group')
        plt.ylabel('Mean')

        plt.show()
```

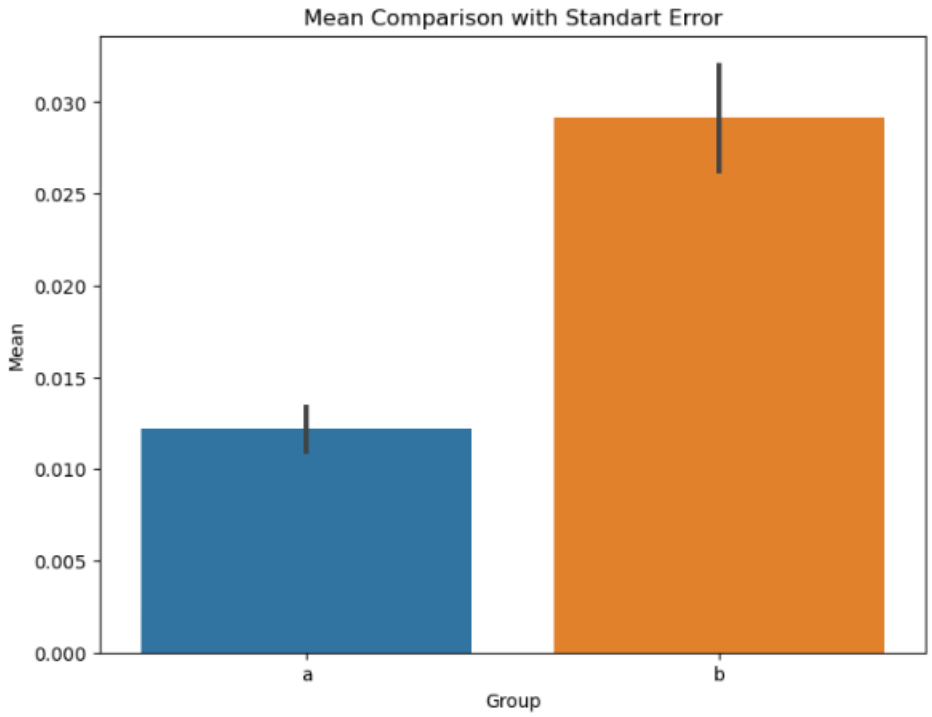


Рис. 2.21 Візуалізація різниці між групами та середньоквадратична похибка

У Python ми можемо візуалізувати різницю між групами та середньоквадратичну похибку, використовуючи функцію seaborn.barplot (Рис. 2.21). Даний графік демонструє похибку цільової метрики. В нашому випадку цільова метрика – це конверсія в покупку.

Крім різниці в значенні цільової метрики, іноді буває також корисно побудувати та порівняти графіки розподілів. Наприклад, для того щоб приблизно оцінити, чи схожий розподіл, який ми отримали, на нормальний та чи можна апроксимувати на нього властивості нормального розподілу.

Така візуалізація буде значно кориснішою у випадку, коли ми розглядаємо неперервні розподіли, тому для цього прикладу ми згенеруємо два нормальні розподіли за допомогою функції stats.norm.rvs.

```
In [6]: import seaborn as sns

        plt.figure(figsize=(10, 6))

        sns.kdeplot(stats.norm.rvs(size=1000))
        sns.kdeplot(stats.norm.rvs(size=1000))

        plt.title('Distribution of A/B Groups')
        plt.xlabel('Value')
        plt.ylabel('Frequency')

        plt.legend(['A', 'B'])
        plt.show()
```

Рис. 2.22 Код на мові Python для візуалізації розподілу

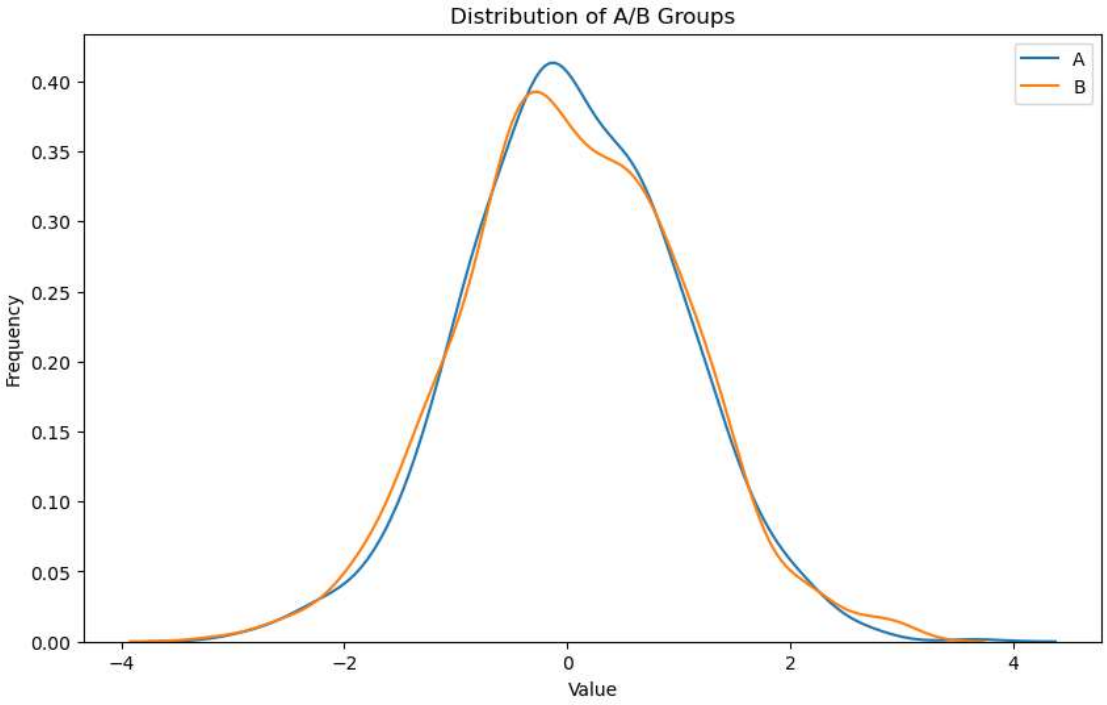


Рис. 2.23 Розподіл а групах A і B

Як бчимо по графікам розподілу (Рис. 2.23), розподіл в обох групах близький до нормального, а це означає, що тестові групи було обрано правильно (по кількості учасників і часу проведення тесту)

Отже, підсумовуючи всі підразунки, можна зробити наступні висновки:

- при тестуванні груп, над якими проводились рекламні кампанії, стало очевидно, що дві різні кампанії спрацювали на різні

категорії користувачів і категорія В виявилась найбільш зацікавленою в покупці. Тому є необхідність дослідити саме цю категорію покупців (вік, гендер, хоббі, рід занять та інше) для розробки наступних кампаній;

- цільова метрика в тестуванні – це конверсія в покупку, вона буде і надалі достатньо високою для подібних кампаній, націлених на певну категорію користувачів.
- слід відхилити нульову гіпотезу (в якій припускалось, що вібірка користувачів А та вибірка В – це одна й та сама цільова аудиторія) та робимо висновок, що різниця між нашими групами А і В є статистично значущою.

ВИСНОВКИ

Отже, для створення інтерактивного дашборду (інформаційної панелі), який буде відображати тенденції в поведінці користувачів інтернет-магазину, потрібно зібрати інформацію про всі кроки користувача сайту. Такі дані може зібрати Google Analytics (якщо користуватись саме інструментами Google). Дані будуть зберігатись у хмарному сховищі. Наступним етапом дані перевіряються на цілісність, наявність дублікатів та пустих значень. Очистка даних може бути реалізована за допомогою мови SQL в Google BigQuery. Далі підготовані дані експортуються в csv-файл. Тут важливо зауважити, що існує немало інструментів, здатних вирішити задачі аналітика на кожному з етапів, але ми використовуємо в рамках завдання саме програмне забезпечення Google.

В Google Looker Studio дані імпортуються напряму або через файл. Саме тут відбувається власне побудова дашборду з графіками, підсумками, фільтрами та висвітленням проблемних і найбільш значущих показників.

Якщо говорити про призначення аналітичних дашбордів, то вони не можуть сказати маркетологу, чи залучати більше людей (рекламуючи товар на Facebook, наприклад), чи робити рекламну кампанію корейською або китайською мовою, чи залишити все як є. Але дашборд здатен показати користувачеві, де є високий результат, а де показники досить низькі, а значить стратегія не працює і варто її змінити.

Важливо також не забувати, що дашборд буде правдивим і інформативним тільки в тому випадку, якщо аналітик правильно збирає та очищує дані, інакше вся подальша робота не тільки не має сенсу, а й може призвести до помилкових рішень в стратегіях компанії.

Результати A/B тестування можуть суттєво допомогти в корегуванні подальшої маркетингової політики компанії.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Адель Світвуд Маркетингова аналітика. Як підкріпити інтуїцію даними: Наш Формат, 2019. – 152 с.
2. Андреас Мюллер та Сара Гвідо Вступ до машинного навчання за допомогою Python. Посібник для фахівців із роботи з даними: O'Reilly Media, 2016. – 398 с.
3. Вес Маккінні Python та аналіз даних: ДМК Прес, 2022. – 552 с.
4. Іцик Бен-Ган Microsoft SQL Server 2012. Основи T-SQL: Microsoft Press, 2012. – 442 с.
5. Карл Андерсон Аналітична культура. Від збору даних до бізнес-результатів: O'Reilly Media, 2017. – 336 с.
6. Ларрі Вассерман All of statistics: Springer, 2010. – 462 с.
7. Майкл Льюїс Moneyball. Як математика змінила найпопулярнішу спортивну лігу у світі: W. W. Norton & Company, 2004. – 336 с.
8. Петер Флах Машинне Навчання: Мистецтво І Наука Алгоритмів, Які Сприймають Дані: Cambridge University Press, 2012. – 416 с.
9. Франсуа Шолле Deep Learning with Python: Manning Publications, 2018. – 384 с.
10. Чарльз Вілан Гола статистика: W. W. Norton & Company 2014. – 282 с.
11. Ясер С. Абу-Мостафа, Малік Магдон-Ісмаїл і Сюан-Тянь Лінь Learning From Data: AMLBook, 2012. – 213 с.
12. DAMA International DAMA-DMBOK: Data Management Body of Knowledge: Technics Publications, 2017. – 588 с.

ДОДАТКИ

Додаток А

		Позначення	Найменування	Кількість аркушів	Примітка
1					
2			Документація		
3					
4		САіУ.КР.22.12 ПЗ	Пояснювальна записка		Формат А4
5					
6		САіУ.РД.13.12 ДМ	Демонстраційні матеріали		Презентація Power Point на компакт диску
7					
8		САіУ.РД.13.12 КР	Копія роботи		(Диск CD-R)

Зм	Лист	№ докум.	Підпис	Дата
Розроб.		Липко С.В.		
К. розд.		Одновол М.М.		
Керівн.		Одновол М.М.		
Н. контр.		Хомяк Т.В.		
Зав. каф.		Желдак Т.А.		

САУ.КР.22.13.ДА.ПЗ

Літ			Аркуш	Аркушів

Матеріали випускної роботи

НТУ «ДП»; 12. 124-20-2