Практична сесія

Аналіз поведінки клієнтів в E-commerce за допомогою Python та Pandas

Частина 1: Завантаження та первинний огляд

1. **Імпортуйте бібліотеку Pandas.**
2. **Завантажте дані з папки work в DataFrame df.**
3. **Проведіть первинний огляд даних:**
   - Виведіть перші 5 рядків (.head()).
   - Перевірте розмір DataFrame (.shape).
   - Вивчіть інформацію про стовпці, їх типи та наявність пропущених значень (.info()).
   - Отримайте базові статистичні показники для числових стовпців (.describe()).
4. **Обговорення в команді:** Дайте відповіді на питання:
   - Чи є в даних пропущені значення?
   - Чи всі стовпці мають коректний тип даних?
   - Які перші спостереження ви можете зробити, дивлячись на .describe() (наприклад, середній вік клієнтів, середні витрати)?

Частина 2: Підготовка даних та створення нових ознак

1. **Створення метрики "Вартість однієї покупки".** Часто важлива не тільки загальна сума витрат, але й середня вартість одного товару.
   - Створіть новий стовпець Avg Spend per Item, поділивши Total Spend на Items Purchased.
   - Перевірте, чи не виникає ділення на нуль. Якщо так, обробіть цей випадок.

2. **Створення сегментів за активністю.** Клієнтів часто ділять на активних та тих, хто давно не робив покупок.
    - Створіть стовпець Activity Status. Якщо Days Since Last Purchase менше 30, присвойте значення 'Active', інакше — 'Inactive'.
    - **Підказка:** Для цього зручно використовувати функцію apply з lambda-виразом.
3. **Перевірка:** Виведіть перші 5 рядків оновленого DataFrame з новими стовпцями, щоб переконатися, що все розраховано правильно.

Частина 3: Дослідницький аналіз

1. **Питання 1: Який тип членства є найціннішим для компанії?**
    - Згрупуйте дані за Membership Type.
    - Для кожного типу розрахуйте середні Total Spend, Items Purchased та Average Rating.
    - Відсортуйте результат за середніми витратами.
    - **Висновок:** У текстовій комірці напишіть, який тип членства приносить найбільше доходу і є найбільш задоволеним.
2. **Питання 2: Як знижки впливають на поведінку клієнтів?**
    - Згрупуйте дані за стовпцем Discount Applied (True/False).
    - Порівняйте середні Total Spend та Items Purchased для клієнтів, які використовували знижку, і тих, хто ні.
    - **Висновок:** Чи дійсно клієнти зі знижками витрачають більше? Чи купують вони більше товарів?
3. **Питання 3: Які вікові групи є найбільш активними та платоспроможними?**
    - Знайдіть 5 наймолодших та 5 найстарших клієнтів у датасеті.
    - Порівняйте їхні середні витрати (Total Spend) та активність (Days Since Last Purchase).
    - **Висновок:** Чи є помітна різниця в купівельній поведінці між різними поколіннями?
4. **Питання 4: Чи дійсно жінки з Gold-членством мають вищий рівень задоволеності, ніж чоловіки з тим же статусом?**
    - Виберіть всіх членів з Gold статусом та згрупуйте їх у по рівні задоволеності (Satisfaction Level) та гендеру (Gender)
    - Розрахуйте відсоток для жінок і для чоловіків і порівняйте їх.
    - **Висновок:** Чи дійсно жінки мають вищий рівень задоволеності, ніж чоловіки?

Частина 4: Візуалізація результатів

Представте ваші знахідки у вигляді графіків.

1. **Графік 1:** Створіть стовпчасту діаграму (bar chart), яка порівнює середні витрати (Total Spend) для кожного Membership Type.
2. **Графік 2:** Створіть діаграму розсіювання (scatter plot), щоб візуалізувати зв'язок між віком (Age) та загальними витратами (Total Spend).
3. **Вимоги:** Кожен графік повинен мати інформативний заголовок та підписані осі.

---

5. Бонусні завдання (додаткові бали)

- **Бонус 1: Розширений аналіз сегментації.**
  - Створіть новий сегмент клієнтів "VIP", до якого належать клієнти з Membership Type = 'Gold' та Total Spend > середнього по всьому датасету.
  - Порівняйте середній рейтинг (Average Rating) та рівень задоволеності (Satisfaction Level) "VIP" клієнтів з усіма іншими.
- **Бонус 2: Аналіз кореляцій.**
  - Створіть кореляційну матрицю для всіх числових стовпців у датасеті.
  - У текстовій комірці вкажіть, які дві змінні мають найсильнішу позитивну кореляцію, і поясніть, що це може означати.