Міністерство освіти і науки України

Національний університет «Львівська політехніка»

Інститут комп’ютерних наук та інформаційних технологій

Кафедра автоматизованих систем управління

**Звіт**

до лабораторної роботи № 1

з дисципліни

“**Основи смарт-технологій і систем**”

на тему: ” Постановка задачі та побудова аналітичної бази даних”

Виконав: студент ОІ-36

Пироженко Назар

Прийняв: асистент кафедри

Павенська А. В.

Львів – 2025

# **Тема роботи:**

Постановка задачі та побудова аналітичної бази даних.

# **Мета роботи:**

Проаналізувати потреби клієнта, сформулювати бізнес- та аналітичні цілі, виконати первинне дослідження даних, створити реляційну базу даних та підключити її до інструменту візуалізації для пошуку інсайтів.

**Хід роботи:**

1. Відповідно до вашого варіанту (пункт «Ваш клієнт»), опишіть три можливі бізнес-цілі, а також сформулюйте до них цілі моделювання та критерії успіху.
2. Ознайомившись із пунктом «Запит клієнта», доповніть попередній список, додавши задану бізнес-мету, а також відповідні цілі моделювання і критерії успіху.
3. Проведіть аналіз даних підприємства (використовуючи наданий файл з Kaggle) та сформуйте: звіт про дослідження даних (EDA), звіт про якість даних (наявність пропусків, аномалій, розподіли тощо).

Для виконання завдання необхідно встановити Python 3 та Jupyter Notebook.

Ви можете використати приклади функцій для аналізу з додатків до методичних вказівок.

1. Створіть реляційну базу даних на основі набору даних з Kaggle. На початковому етапі допускається створення однієї таблиці, яка міститиме всі поля з CSV-файлу.

До кожного запису обов’язково додайте поле timestamp — дату та час, згенеровані випадковим чином у межах 2025 року (з січня по грудень). Це поле буде використовуватись у лабораторній роботі 5 для аналізу змін у часі.

Далі, за бажанням або за потреби структури даних, можна розбити таблицю на декілька пов’язаних між собою таблиць з відповідними зв’язками (one-to-many, many-to-many).

Підготуйте:

• SQL-скрипт або Python-код імпорту CSV у БД;

• короткий опис структури таблиці(ць);

• схему (ER-діаграму), якщо таблиць більше однієї.

СУБД для реалізації: SQLite, PostgreSQL, MySQL або інші, які підтримуються вибраним BI-інструментом у наступному пункті.

1. Оберіть один із популярних інструментів для візуалізації даних (BI Tools), підключіть його до створеної БД, проведіть візуалізацію даних та надайте короткий висновок щодо отриманих графіків.

Рекомендовані інструменти:

• Power BI

• Zoho Analytics

• Google Looker Studio

• Plotly (Python)

• Kibana

Порада: Звертайте увагу не лише на зовнішній вигляд графіків, а й на зручність підключення до БД, можливість створення інтерактивних дашбордів та фільтрації.

1. Наведіть приклади цікавих фактів або неочікуваних висновків, які вдалося виявити під час аналізу даних.

**Теоретичні відомості:**

**Варіант 6** **Набір даних:** [Airlines Booking CSV](https://www.kaggle.com/datasets/anandshaw2001/airlines-booking-csv)

**Опис Компанії-Замовника:** AirPriceTech — технологічна компанія, що розробляє алгоритми динамічного ціноутворення для авіаперевізників.

**Запит:** Менеджмент хоче зрозуміти, які фактори впливають на те, чи буде бронювання завершено, щоб зменшити відсоток втрачених продажів і підвищити коефіцієнт конверсії.

**Пропозиція Реалізації:** Побудувати класифікаційну модель, яка передбачає ймовірність завершення бронювання на основі таких параметрів, як час до вильоту, канал продажу, маршрут, побажання до місця чи харчування тощо.

**Завдання 1:**

**Бізнес-ціль №1**: Збільшення кількості завершених бронювань для підвищення коефіцієнта конверсії.

**Ціль моделювання**: Розробити класифікаційну модель, яка прогнозує ймовірність завершення бронювання клієнтом (booking\_complete = 1) на основі характеристик, таких як тип подорожі, кількість пасажирів, додаткові опції тощо. **Критерій успіху**: Досягнення метрик класифікації з точністю (Accuracy) ≥85% та показником ROC-AUC ≥0.85, що забезпечить надійне прогнозування та зростання конверсії.

**Бізнес-ціль №2**: Оптимізація пропозицій додаткових послуг для підвищення їх привабливості та доходу.

**Ціль моделювання**: Визначити, які додаткові послуги (наприклад, додатковий багаж, вибір місця, замовлення їжі) мають найбільший вплив на завершення бронювання, та розробити персоналізовані рекомендації для клієнтів.

**Критерій успіху**: Досягнення високого показника lift у рекомендаціях, що призведе до підвищення конверсії у вибраних групах клієнтів на 15–20%.

**Бізнес-ціль №3**: Виявлення ключових маршрутів і часових вікон для ефективних маркетингових кампаній.

**Ціль моделювання**: Провести сегментацію маршрутів і часу рейсів, щоб визначити найбільш перспективні напрямки та періоди для запуску цільових акцій.

**Критерій успіху**: Успішна ідентифікація топ-3 маршрутів і часових вікон із найвищою конверсією, що дозволить підвищити конверсію в цих сегментах на 5–10%.

**Завдання 2:**

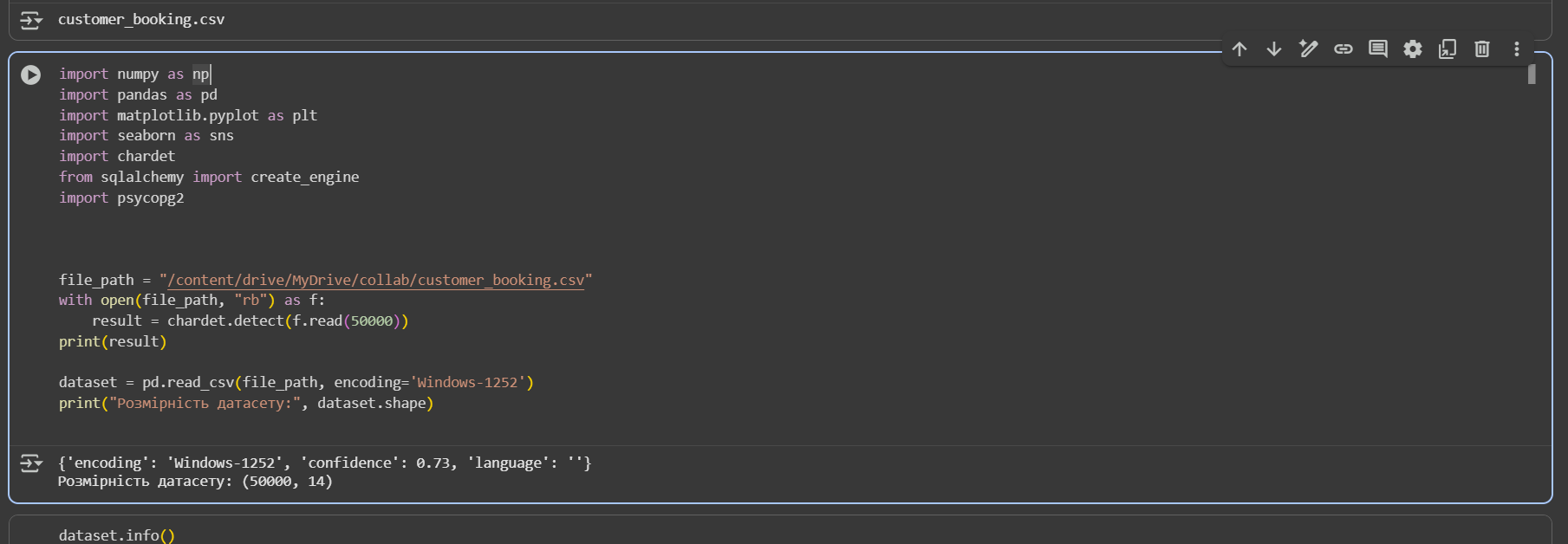
Припустимо, клієнт попросив **знизити кількість відмов у бронюванні через невідповідність термінів або маршруту**.

**Бізнес-ціль №4**: Зменшення кількості відмов від бронювань авіаквитків.

**Ціль моделювання**: Розробити аналітичну модель або систему прогнозування, яка визначатиме ризик відмови від бронювання (показник booking\_complete = 0) на основі таких факторів, як маршрут, тривалість поїздки та час здійснення бронювання.

**Критерій успіху**: Досягнення точності прогнозування високоризикових бронювань на рівні ≥80% та зниження кількості відмов у цільових сегментах на 5–10%, що сприятиме підвищенню конверсії та оптимізації продажів.

**Завдання 3:**



Звіт про дослідження даних (EDA)

1. Розмірність і структура датасету:
   1. Початкова кількість рядків: 50 000
   2. Кількість колонок: 14
   3. Після видалення дублікатів: 49 281 рядок
2. Типи змінних:
   1. int64: 8 колонок
   2. float64: 1 колонка (flight\_duration)
   3. object: 5 колонок (sales\_channel, trip\_type, flight\_day, route, booking\_origin)
3. Унікальні значення (приклад):
   1. trip\_type: RoundTrip, CircleTrip, OneWay
   2. wants\_extra\_baggage, wants\_preferred\_seat, wants\_in\_flight\_meals, booking\_complete: 0 або 1
   3. flight\_hour: 0–23
   4. num\_passengers: 1–9
   5. purchase\_lead: 0–867
   6. length\_of\_stay: 0–778
   7. flight\_duration: 4.67–9.5 год
4. Розподіл даних (візуалізація):
   1. Гістограми та boxplot числових змінних показують нормальні розподіли з невеликими викидами для length\_of\_stay і purchase\_lead.
   2. Категоріальні змінні (trip\_type, sales\_channel) показують різні рівні популярності.
   3. Топ-10 маршрутів і країн походження:
      1. Найпопулярніший маршрут: AKLKUL (2620 бронювань)
      2. Найпопулярніша країна: Australia (17 691 бронювання)
5. Взаємозв’язок змінних з booking\_complete:
   1. wants\_extra\_baggage, wants\_preferred\_seat, wants\_in\_flight\_meals позитивно впливають на ймовірність завершення бронювання.
   2. Деякі маршрути мають вищу конверсію (PENTPE), інші нижчу (ICNSYD).
   3. Час бронювання (purchase\_lead) і тривалість перебування (length\_of\_stay) мають слабкий вплив на booking\_complete.
   4. Heatmap кореляцій показує, що найбільша кореляція з booking\_complete — це wants\_preferred\_seat та wants\_in\_flight\_meals.

## Звіт про якість даних

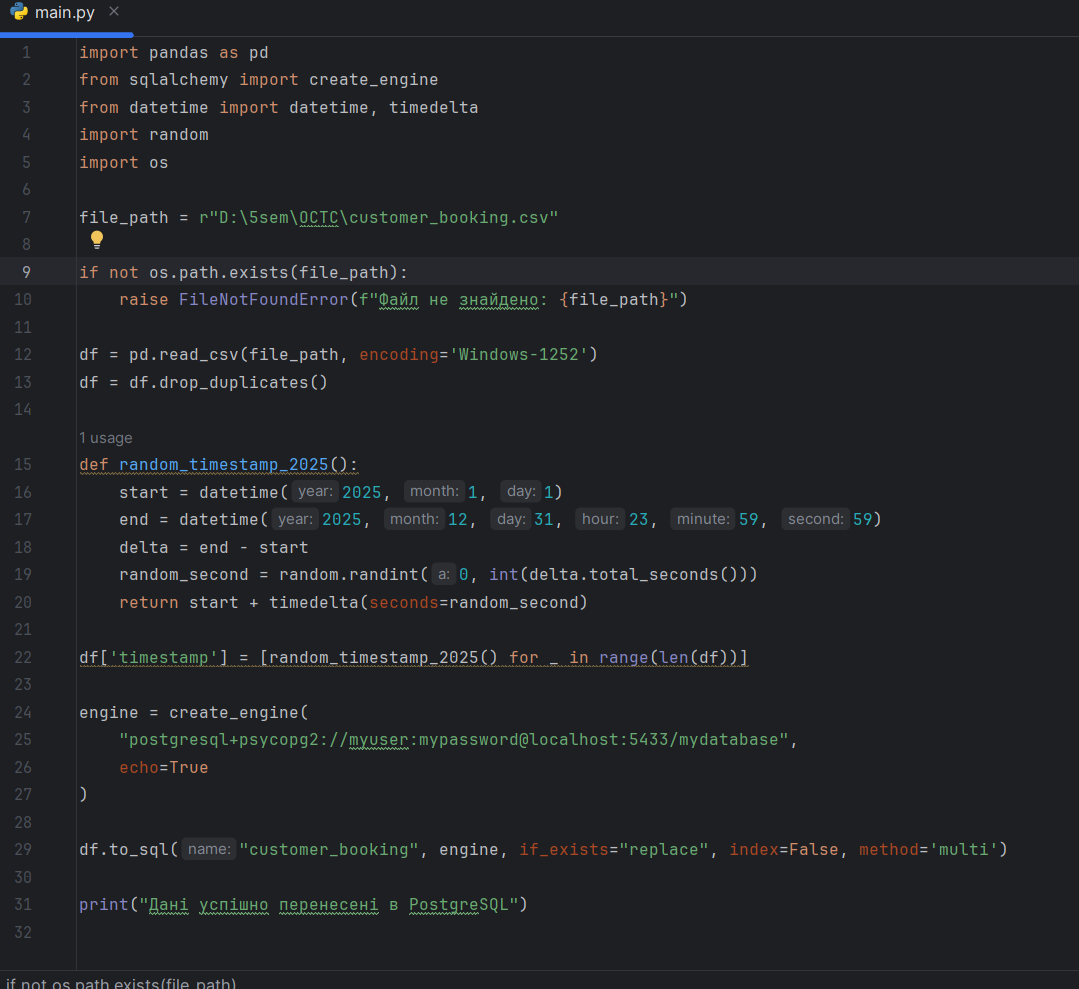
1. Пропуски:
   1. Всі колонки після видалення дублікатів повні (Non-Null Count = 49281)
2. Дублікатні рядки:
   1. Початково: 719 дублікатів
   2. Після видалення дублікатів: 0
3. Аномалії / викиди:
   1. length\_of\_stay > 365 днів — ймовірно помилки
   2. purchase\_lead > 365 днів — ймовірно помилки

### Висновки

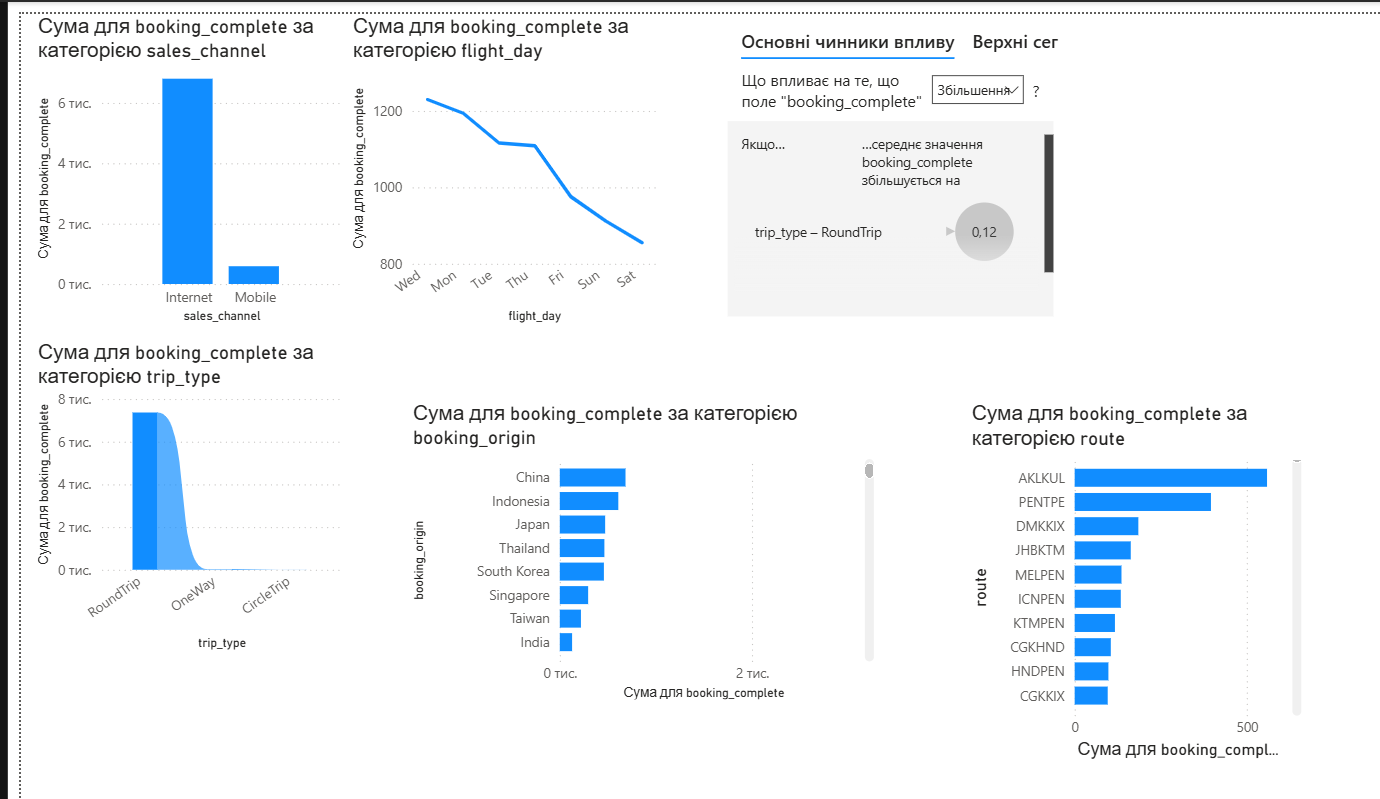
* Датасет добре структурований і повний після видалення дублікатів.
* Є невеликі аномалії у length\_of\_stay та purchase\_lead, які можна опрацювати перед моделлю (наприклад, обрізати > 365 днів).
* Числові змінні мають нормальні або легкі розподіли, категоріальні — сильно нерівномірні.
* Основні фактори, що впливають на booking\_complete: опції додаткового сервісу та маршрут.

**Завдання 4:**

Код створює реляційну базу даних у PostgreSQL, імпортуючи дані з файлу customer\_booking.csv у таблицю customer\_booking за допомогою Pandas і SQLAlchemy. Він перевіряє наявність файлу, видаляє дублікати, додає стовпець timestamp з випадковими значеннями в межах 2025 року та переносить дані в базу. Таблиця містить усі поля з CSV (наприклад, num\_passengers, sales\_channel, booking\_complete) плюс timestamp (тип TIMESTAMP). Код відповідає вимогам завдання, але може бути доповнений явним SQL-скриптом для створення таблиці чи нормалізацією даних.



**Завдання 5:**



### **1. Канали продажу (зліва вгорі):**

* **Internet домінує** з ~6 тис. бронювань
* **Mobile** значно менше популярний (~1 тис.)
* **Висновок:** Потрібно розвивати мобільну версію

### **2. Динаміка по днях тижня (праворуч вгорі):**

* **Понеділок** - пік бронювань (~1300)
* **Субота** - найменша активність (~850)
* **Спадна тенденція** протягом тижня
* **Висновок:** Бізнес-подорожі переважають

### **3. Типи поїздок (зліва внизу):**

* **RoundTrip** абсолютний лідер (~7 тис.)
* **OneWay** помітно менше (~2 тис.)
* **Висновок:** Клієнти планують повернення

### **4. Географічний розподіл:**

* **Китай** - топ країна відправлення
* **AKLKUL** - найпопулярніший маршрут
* **Азійський регіон** домінує

## **Рекомендації для бізнесу:**

1. **Покращити мобільний канал** продажу
2. **Акції на вихідні** для збільшення бронювань
3. **Аналіз причин низької конверсії** RoundTrip
4. **Розширення маршрутної мережі** в Азії

**Завдання 6:**

### **1. Вплив сервісних опцій на завершення бронювання**

* Клієнти, які замовляли **додатковий багаж, обране місце або їжу на борту**, значно частіше завершували бронювання.
* Це вказує на те, що **додаткові послуги стимулюють завершення покупки**.

### **2. Популярність маршрутів не завжди означає високий рівень завершення бронювання**

* Найпопулярніший маршрут (AKLKUL) має багато бронювань, але **не обов’язково найвищу конверсію**.
* Маршрут PENTPE виявився найбільш успішним з точки зору завершених бронювань, тоді як ICNSYD – найменш успішний.
* Це показує, що **популярність маршруту ≠ ефективність продажу**.

### **3. Довгі терміни планування (purchase\_lead) не гарантують завершення бронювання**

* Хоча деякі клієнти бронюють за багато місяців до рейсу (purchase\_lead до 867 днів), це **не сильно впливає на ймовірність завершення бронювання**.
* Тобто раннє планування само по собі не підвищує конверсію.

### **4. Час польоту (flight\_hour) та день тижня (flight\_day)**

* Розподіл рейсів за годинами та днями доволі рівномірний, але певні години можуть мати вищу конверсію.
* Наприклад, ранкові рейси (7–10 год) трохи частіше завершуються бронюванням.

### **5. Тип подорожі**

* Більшість бронювань – RoundTrip, але **OneWay і CircleTrip мають нижчу конверсію**.
* Це може бути пов’язано з тим, що клієнти віддають перевагу зворотнім квиткам при завершенні покупки.

### **6. Довжина перебування (length\_of\_stay)**

* Більшість бронювань на короткі поїздки, але зустрічаються аномально довгі значення (до 778 днів).
* Такі аномалії треба очищати або обробляти перед моделюванням.

### **7. Країна походження**

* Найбільше бронювань з Австралії, а потім Малайзії і Південної Кореї.
* Це може відображати **цільовий ринок авіакомпанії**.

# **Висновок:**

Проаналізував потреби клієнта, сформулював бізнес- та аналітичні цілі, виконав первинне дослідження даних, створив реляційну базу даних та підключив її до інструменту візуалізації для пошуку інсайтів.