📂 DATA MAAN LO IS TYPE KA HAI

Columns (example):

- PassengerID
- Name
- Age
- Gender
- TrainNo
- TrainName
- Source
- Destination
- Class (SL / 3A / 2A / 1A)
- TicketPrice
- BookingDate
- TravelDate
- SeatNo
- Coach
- PaymentMode
- BookingStatus (Confirmed / WL / Cancelled)

🧠 TUMHARA KAAM KYA HOGA? (STEP-BY-STEP TASK LIST)

1️⃣ DATA UNDERSTANDING (Sabse pehle)

👉 Ye sirf **dekhne aur samajhne ka step** hai

- Total passengers count

- Kitne columns hain
- Kaunse column numeric / text / date
- Missing values hain ya nahi

📌 **Output expectation:**

“Dataset me 512 passengers aur 16 columns hain.”

2️⃣ DATA CLEANING (Bahut important ❗)

◇ Missing values

- Age missing?
- SeatNo missing?
- PaymentMode blank?

📌 Decide:

- Fill karna hai?
- Ya row delete karni hai?

◇ Duplicate passengers

- Same PassengerID repeat ho raha hai?
- Same name + train + date repeat?

📌 Duplicate remove karo

◇ Data type correction

- Date columns → date format
- TicketPrice → numeric
- Age → integer

◇ Text cleaning

- City name uniform karo (Delhi / delhi / DELHI ❌)
- Gender values fix (M / Male / male ❌)

3️⃣ BASIC ANALYSIS (Warm-up 🔥)

📊 Passenger Overview

- Total passengers
- Male vs Female ratio
- Average age
- Min / Max age

4️⃣ TRAIN & ROUTE ANALYSIS (Real Insight)

🚆 Train-wise

- Kaunsa train sabse zyada passengers le ja raha hai
- Kaunsa train sabse mehnga hai (avg ticket)

🌍 Route-wise

- Top source cities
- Top destination cities
- Most traveled route (Source → Destination)

5️⃣ CLASS ANALYSIS (Revenue ka heart ❤️)

- SL / 3A / 2A / 1A distribution
- Kaunsi class se sabse zyada revenue
- Avg ticket price per class

📌 **Insight likho:**

“3A class generates 42% of total revenue.”

6️⃣ BOOKING & PAYMENT ANALYSIS

💳 PaymentMode

- UPI / Card / Cash / NetBanking count
- Sabse zyada used payment method

🎟️ BookingStatus

- Confirmed vs Waiting vs Cancelled
- Cancellation rate (%)

7️⃣ DATE & TIME ANALYSIS

📅 BookingDate vs TravelDate

- Advance booking days calculate karo
- Most bookings kis month me hui

⏰ Peak time insights

- Festival season bookings
- Weekend vs weekday travel

8️⃣ REVENUE ANALYSIS (MOST IMPORTANT 💰)

- Total revenue
- Train-wise revenue
- Route-wise revenue
- Class-wise revenue

📌 **Business-style statement likho:**

“Top 3 trains contribute 55% of total revenue.”

9️⃣ OUTLIERS & ANOMALIES

- Extremely high ticket prices
- Age < 5 ya Age > 90
- One passenger booking too many tickets

👉 Ye **fraud / error detection** hota hai (industry use)

🔟 DATA VISUALIZATION (Decision ke liye)

📊 Banana hai:

- Bar chart → Train-wise passengers
- Pie chart → Class distribution

- Line chart → Monthly bookings
- Bar chart → Revenue by class

1️⃣1️⃣ FINAL INSIGHTS & STORY (MOST CRITICAL 🔥)

❌ Sirf number nahi
✅ **Meaning likho**

Examples:

- “Delhi–Mumbai route is the busiest with highest revenue.”
- “Sleeper class has highest passengers but lower revenue per passenger.”
- “UPI is the most preferred payment mode.”

1️⃣2️⃣ FINAL EXPORT / REPORT

- Cleaned CSV
- Summary CSV (train-wise, route-wise)
- Graphs
- 1–2 page insight summary (PDF / doc)

📊 PART 2️⃣ — DATA ANALYSIS QUESTIONS (Step-by-Step Tasks)

Is CSV par **sirf ye kaam karo** 👇
(ye exactly wahi cheez hai jo companies expect karti hain)

🔹 A. Data Understanding

1. Total passengers kitne hain?
2. Dataset me kitne columns hain?
3. Kaunse columns numeric, text aur date type ke hain?
4. Missing values ka count column-wise.

🔹 B. Data Cleaning

5. Missing values ko handle karo (fill / drop).
6. Duplicate passengers identify karo.
7. BookingDate aur TravelDate ko datetime me convert karo.
8. City names aur Gender values ko standard format me lao.

🔹 C. Basic Passenger Analysis

9. Average age of passengers.
10. Male vs Female passenger count.
11. Minimum aur maximum age.

◇ D. Train & Route Analysis

12. Kaunsa train sabse zyada passengers le ja raha hai?
13. Top 5 source cities.
14. Top 5 destination cities.
15. Sabse zyada travel hone wala route (Source → Destination).

◇ E. Class Analysis (IMPORTANT 💰)

16. Class-wise passenger count.
17. Class-wise average ticket price.
18. Kaunsi class sabse zyada revenue generate karti hai?

◇ F. Revenue Analysis

19. Total revenue calculate karo.
20. Train-wise total revenue.
21. Route-wise revenue.
22. Top 3 revenue generating trains.

◇ G. Booking & Payment Analysis

23. PaymentMode ka distribution.
24. Sabse popular payment method.
25. Confirmed vs Cancelled booking percentage.
26. Cancellation rate calculate karo.

◇ H. Date-Based Analysis

27. Booking aur travel ke beech average days.
28. Month-wise booking trend.
29. Peak booking month identify karo.

◇ I. Outlier / Anomaly Detection

30. Bahut zyada ticket price wale records.
31. Age outliers (Age < 5 ya Age > 75).

◇ J. Visualization (Minimum graphs)

32. Bar chart → Train-wise passengers.
33. Pie chart → Class distribution.
34. Line chart → Monthly bookings.
35. Bar chart → Revenue by class.

📝 PART 3️⃣ — **FINAL PROJECT REPORT FORMAT (Copy–Paste Ready)**

📌 1. Project Title

Train Passenger Data Analysis using Python (Pandas)

📌 2. Objective

Is project ka objective train passenger data ko clean, analyze aur insights nikalna hai jisse:

- Passenger behavior samjha ja sake
- Revenue patterns identify ho

- Popular routes aur trains ka pata chale

📌 3. Dataset Description

Dataset me 520 passengers ka data hai jisme age, gender, route, class, ticket price, booking date, payment mode aur booking status shamil hai.

📌 4. Tools Used

- Python
- Pandas
- Matplotlib

📌 5. Data Cleaning Steps

- Missing values handle kiye
- Duplicate records remove kiye
- Date columns ko datetime format me convert kiya
- Text data ko standard format me normalize kiya

📌 6. Analysis Performed

- Passenger demographics analysis
- Train-wise aur route-wise analysis
- Class-wise revenue analysis
- Payment mode aur booking status analysis
- Date-based trend analysis

📌 7. Key Insights

(Example – tum apne output ke hisab se likhna)

- Sleeper class me sabse zyada passengers hain lekin revenue AC classes se zyada aata hai
- Delhi aur Mumbai sabse busy source cities hain
- UPI sabse preferred payment method hai
- Top 3 trains total revenue ka major hissa generate karte hain

📌 8. Conclusion

Is project se clear hota hai ki passenger data analysis se operational aur business decisions improve kiye ja sakte hain. Pandas ka use karke large datasets ko efficiently analyze kiya ja sakta hai.