DATA ANALYSTS INTERNSHIP

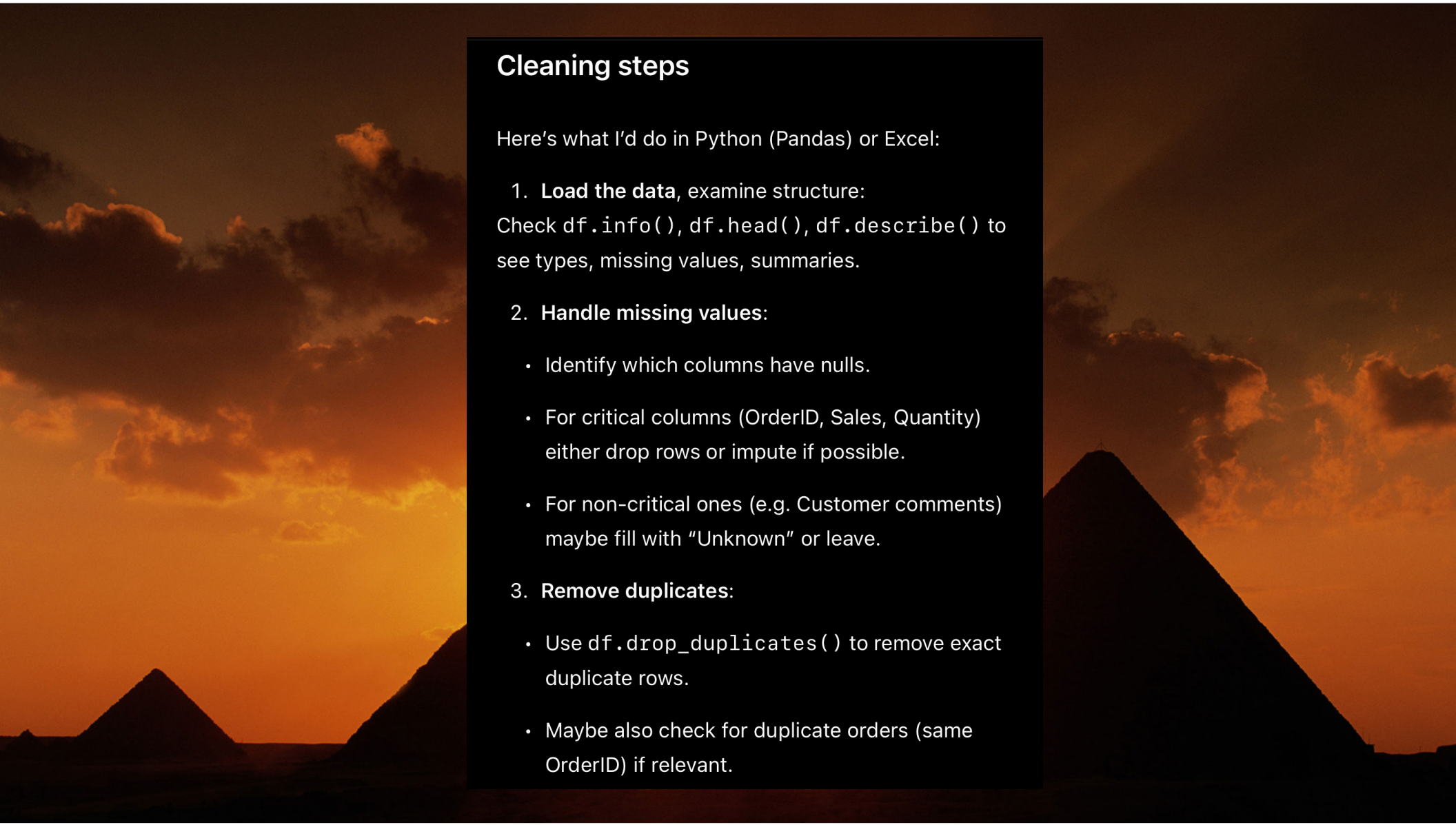
Here's a plan & sample of how I would clean a "Sales Data" dataset. Since I don't have a specific file from you, I'll take the **Superstore Sales** dataset from Kaggle as a working example. Kaggle +1

I'll outline what steps I'd do, show sample before-after of what is cleaned, and you can apply same to your dataset (or send me your file so I directly clean it for you).

Example Dataset: Superstore Sales

What it contains (before cleaning):

- Orders data: order dates, ship dates, order IDs, customer info, product category, sub-category, sales, profit, quantity, discount, region, etc. Kaggle
- Data types might be mixed (dates stored as text, numeric fields as strings, etc.)
- May include duplicates, missing values, inconsistent formatting in text fields (e.g. “Furniture” vs “furniture” etc.)



Cleaning steps

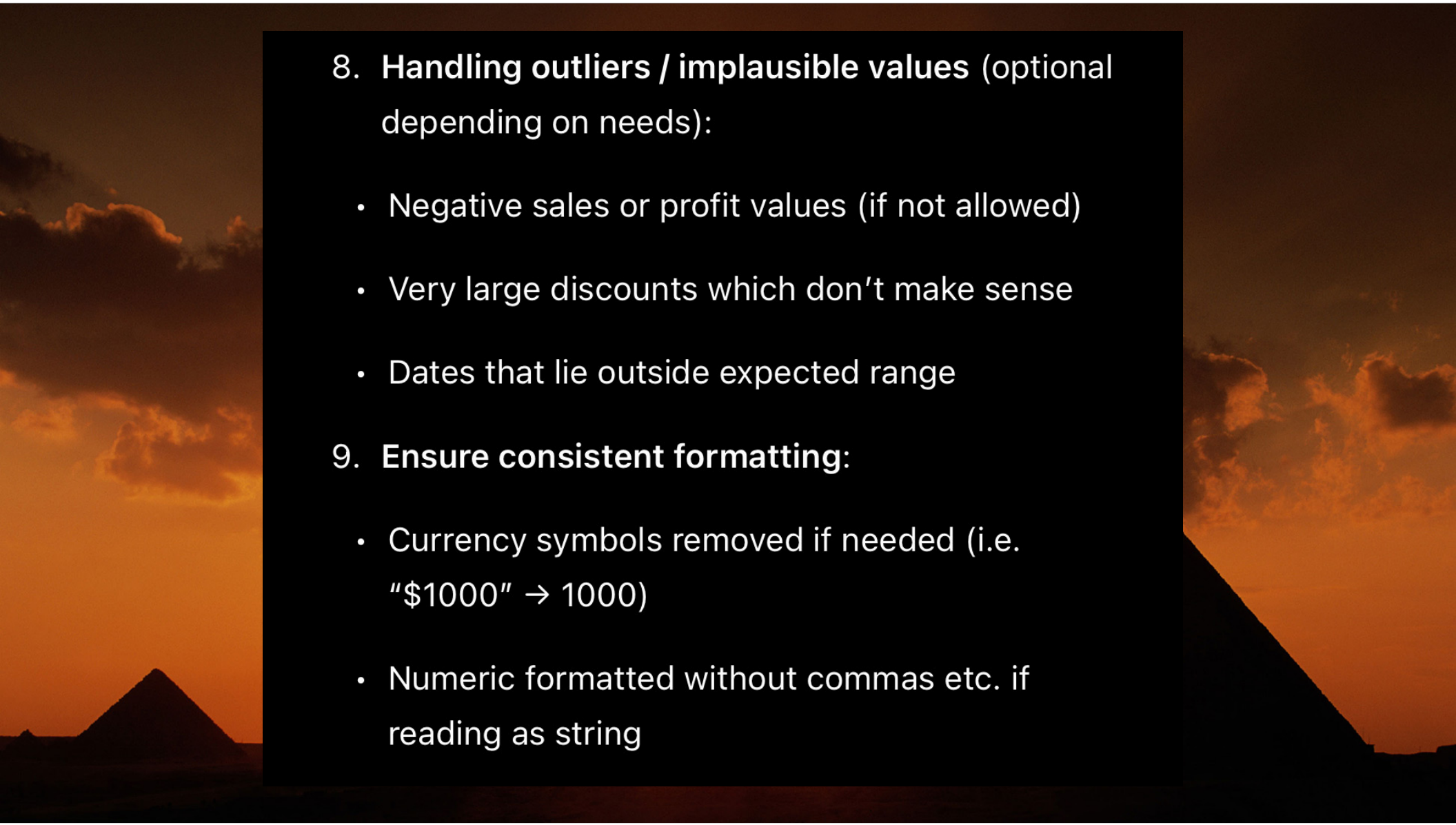
Here's what I'd do in Python (Pandas) or Excel:

1. **Load the data**, examine structure:
Check `df.info()`, `df.head()`, `df.describe()` to see types, missing values, summaries.

2. **Handle missing values**:
   - Identify which columns have nulls.
   - For critical columns (OrderID, Sales, Quantity) either drop rows or impute if possible.
   - For non-critical ones (e.g. Customer comments) maybe fill with "Unknown" or leave.

3. **Remove duplicates**:
   - Use `df.drop_duplicates()` to remove exact duplicate rows.
   - Maybe also check for duplicate orders (same OrderID) if relevant.

4. **Standardize text fields**:
   - Make all text fields consistent: e.g. Category, Sub-Category, Region, Customer Segment etc.
   - Lowercase or Title case as needed; remove leading/trailing whitespace.
5. **Fix date formats**:
   - Convert OrderDate, ShipDate columns to `datetime` dtype.
   - Ensure consistent format (e.g. dd-mm-yyyy or yyyy-mm-dd).
6. **Fix data types**:
   - Numeric fields (Sales, Profit, Quantity, Discount) should be floats/ints.
   - Boolean / categorical fields where possible.
7. **Rename columns** (if needed) for consistency:
   - E.g. “Order Date” → “order_date”, “Ship Mode” → “ship_mode”, remove spaces, all lowercases or underscores.



8. **Handling outliers / implausible values** (optional depending on needs):
   - Negative sales or profit values (if not allowed)
   - Very large discounts which don't make sense
   - Dates that lie outside expected range
9. **Ensure consistent formatting**:
   - Currency symbols removed if needed (i.e. "$1000" → 1000)
   - Numeric formatted without commas etc. if reading as string

Field	Before Cleaning	After Cleaning
OrderDate	“1/3/2014” or “01-Mar-2014” mixed, stored as text	Converted to datetime, all in format YYYY-MM-DD
Sales	Some entries as “$1,000”, some as “1000”, maybe string dtype	All numeric (float), no currency symbol or comma
Category	“Furniture”, “ furniture“, “furniture ” (leading/trailing spaces), mixed case	Cleaned to “Furniture” (title case), no trailing spaces
Duplicates	Some exact duplicate rows	Removed duplicates, kept first occurrence
Nulls in Discount or Profit	Some rows missing discount or profit	Impute (maybe 0 for missing discount) or mark “Unknown”, depending on business logic

What a Cleaned Version Looks Like (structure)

After cleaning, the dataset might have columns like:

- order_id (string)
- order_date (datetime)
- ship_date (datetime)
- ship_mode (category)
- customer_id (string)
- segment (category)
- country / region (category)
- product_category (category)
- product_subcategory (category)
- product_name (string)
- sales (float)
- quantity (int)
- discount (float)
- profit (float)

THANKYOU

Yours Faithfully
Madhvi Lakhotia