🎯 Presentation Title:

🧼 Data Cleaning Analysis

1. Title Slide

Title: Data Cleaning Analysis
Subtitle: Ensuring Data Quality for Reliable Insights
Presented by: Faleye Doyin Opeyemi
Date: 12-8-2025

2. Why Data Cleaning Matters

- Dirty data leads to misleading insights
- Clean data improves model accuracy and decision-making
- This training covers 5 key cleaning concepts

3. Modules Overview

Module	**Topic**	**Goal**
1	Data Integrity	Ensure logical consistency
2	Missing Data	Handle gaps in the dataset
3	Duplicate Removal	Eliminate redundant records
4	Standardization	Harmonize formats and labels
5	Outlier Detection	Manage extreme values

4. Module 1 – Data Integrity

- Import Labries
- Import Dataset

- Check data types and logical rules
- df.info()
- df.describe()

🟦 Slide 5: Module 2 – Missing Data Handling

- Checking for missing values
- Check the column data
- Drop columns missing critical fields
- Check for how many rows and columns are in the dataset

🟦 Slide 6: Module 3 – Duplicate Removal

- Why it matters: Duplicates distort analysis
- Check for duplicates
- df.drop_duplicates(inplace=True)

🟦 Slide 7: Module 4 – Standardization

- Standardize:
    - Using (text col) for all the text column category to standardize all the text columns to lower cases and remove extra spaces.
- Example:
- df['textcol'] = df['textcol'].str.lower().str.strip()

8. Module 5 – Outlier Detection

- Why it matters: Outliers skew results
- 🛠 Techniques:
    - IQR
    - Boxplots – to detect outliers before and after the outlier calculation

- 🧪 Example:
- Q1 = df['price'].quantile(0.25)
- Q1 = df['price'].quantile(0.25)
- IQR = Q3-Q1
- threshold = 1.5 -------# threshold formular
- upperbound and lowerbound formular after the outlier calculation
  lowerbound = Q1 - threshold * IQR
  upperbound = Q3 + threshold * IQR
- Remove outlier from df price
  df = df[(df['price'] >= lowerbound) & (df['price'] <= upperbound)]
- **df.shape** - checking for the accuracy of the rows and columns after the Outliers
- **Visualizing:** with Box plot to re-detect after the removal of outlier

9. Final Cleaning Checklist

Before analysis:

- Correct data types
- Missing values handled
- Duplicates removed
- Formats standardized
- Outliers addressed

10. Pro Tips

- Use pandas-profiling for quick audits
- Automate cleaning steps
- Document every decision

11. Q&A

Any questions or clarifications?
Let’s discuss real-world examples or challenges you’ve faced.

- **Let’s discuss your feedback, or ideas for next steps.**
- **Contact info**: (08130227444)
- **Email:** adesuwadoyinsola@gmail.com