Reading — Stage 08: Exploratory Data Analysis (EDA)

EDA is the investigative phase where you **learn the data's story**. It is not a checkbox—its insights shape the entire modeling process.

Why EDA?

- Reveals distributional shape (skew, tails) that affects model validity.
- Uncovers outliers and data quality issues.
- Surfaces relationships that suggest **feature ideas** and **model forms**.

A Minimal EDA Checklist

1. **Structure & Missingness**
    - `df.info()`, `df.isna().sum()`, datatype expectations
2. **Numeric Profile**
    - `df.describe()` plus `skew`, `kurtosis`
3. **Distributions**
    - Histograms/KDE + boxplots (outliers)
4. **Relationships**
    - Scatter/line; color by category to reveal clusters
5. **Correlation (Carefully)**
    - Use as a hint; not a causal claim

Interpretation Patterns

- **Right-skewed targets** → log transforms before linear modeling
- **Outliers** → winsorize or robust estimators
- **Heteroskedasticity** → consider variance-stabilizing transforms
- **Seasonality/trends** → calendar/time features later

Communication

- Every plot gets a plain-English caption: *What? So what? Now what?*
- Conclude with **3 insights + assumptions** to guide preprocessing & feature engineering.

Example Diagram (suggested slide)

```
Raw Data → Profiling → Visual Patterns → Insights → Assumptions → Next Steps
(Cleaning/Features)
```