1. Why check assumptions in linear regression?

Answer:
Because linear regression makes strong assumptions about the data — like **linearity**, **no multicollinearity**, **homoscedasticity**, **normality of residuals**, and **independent errors**.
If these assumptions are violated, the model's **coefficients**, **interpretability**, and **predictions** may become unreliable.

---

2. Why Shapiro-Wilk Test?

Checks if residuals are **normally distributed**.

- Needed when you want to **interpret coefficients** or **compute p-values**.
- If violated → linear regression may still predict, but statistical inferences become **invalid**.

---

3. Why Breusch-Pagan / White's Test?

Checks for **heteroscedasticity** — whether residuals have **constant variance**.

- If variance changes across predictions → model is unstable.
- Helps decide if you need **robust standard errors** or a **different model**.

---

4. Why Durbin-Watson Test?

Checks for **autocorrelation** in residuals.

- Especially important in **time-series** or ordered data.
- If violated, predictions can be biased.

---

5. Why Ramsey RESET Test?

Checks **model specification** — detects if important features or non-linear terms are missing.

- If failed → linear regression is underfitting; we may need to add **interaction** or **polynomial terms**.

---

6. Why VIF, Condition Number, and Tolerance?

These detect **multicollinearity** between predictors:

- **High VIF (>10)** → features are strongly correlated → coefficients become unreliable.
- Guides you to **drop**, **combine**, or **transform** features.

---

7. Why Mutual Information & Encoding?

- **Mutual Information** → selects **most informative features** for better model performance.
- **Label/One-Hot Encoding** → converts categorical variables into numeric form for modeling.