

UNIVARIATE ANALYSIS REPORT

1. Pregnancies

- Distribution: Right-skewed
- Insight: Most women have 0–3 pregnancies; few high values.
- Outliers: Present
- Interpretation: Expected for count data.

2. Glucose

- Distribution: Nearly symmetric
- Insight: Values cluster around 100–140.
- Outliers: Some high values.
- Interpretation: Reliable predictor.

3. Blood Pressure

- Distribution: Slight left-skew
- Insight: Contains impossible zeros.
- Interpretation: Zeros indicate missing data.

4. Skin Thickness

- Distribution: Slight right-skew
- Insight: Many zeros.
- Interpretation: Should be imputed.

5. Insulin

- Distribution: Heavily right-skewed
- Insight: Many zeros and extreme highs.
- Interpretation: Needs cleaning or transformation.

6. BMI

- Distribution: Slight right-skew
- Insight: Mostly 25–35.
- Interpretation: Reliable feature.

7. Diabetes Pedigree Function

- Distribution: Right-skewed
- Insight: Most values are low; some high-risk outliers.
- Interpretation: Scaling recommended.

8. Age

- Distribution: Right-skewed
- Insight: Most between 20–40.
- Interpretation: Expected population pattern.

9. Outcome

- Distribution: Imbalanced binary
- Insight: More non-diabetic cases.
- Interpretation: May require class balancing.

OVERALL SUMMARY

- Features with missing/impossible values: BloodPressure, SkinThickness, Insulin
- Right-skewed features: Pregnancies, SkinThickness, Insulin, DPF, Age
- Strong outliers: Insulin, SkinThickness, BloodPressure
- Good distributions: Glucose, BMI
- Class imbalance exists and must be handled in modeling.