

Nama: Komang Andika Wira Santosa

NIM: 2501994424

1. Why is the Gaussian Assumption important in EDA?

Numerical data often has imprecision or inaccuracy, which can be handle by using Gaussian Assumption. In EDA, the Gaussian Assumption is important because it allows us to identify outliers using three-sigma edit rule and characterize the distribution of the data using Gaussian probability density function. Gaussian Assumption is useful when the data conforms to a Gaussian distribution and helps us to detect outliers and skewness in the data, which is important in EDA.

2. How can we tell whether the Gaussian assumption is reasonable or not using QQ-Plot?

If the points fall within the straight line, the data is reasonable for Gaussian assumption. If the points deviate significantly from the straight line, it may suggest that the data is not normally distributed and thus not reasonable for Gaussian assumption.

3. How can we tell whether the Gaussian assumption is reasonable or not using Histogram?

If the histogram we created is shaped like a bell and the shape of the histogram appear like the parametric Gaussian density curve, it may suggest that the data is normally distributed and reasonable for Gaussian assumption.

4. Which tool is better to tell the reasonableness of the Gaussian assumption for the data, QQ-plot or Histogram?

The usage of QQ-plot or Histogram have each of their advantages. QQ-plot is better at detecting skewness in the tails of distribution from our data because when the data has deviations from normality, the data will deviate

from the straight line and in histogram is harder if the data is not well centered. Histogram is better for detecting deviations in the middle because the frequencies of intervals used to detect skewness between or in the middle of the distribution. In my opinion, QQ-plot is better because it works better in most cases.

5. What are the strengths and weaknesses of Hampel Identifiers, the three-sigma edit rule, and the boxplot outlier rule in detecting univariate outliers?

#### Three-Sigma Edit Rule

- + Easy to apply.
- + Well known.
- Often performs poorly

#### The Hampel Identifier

- + Better to detect extreme outlier.
- + Can be applied to several distribution.
- + More resistant to the influence of outliers.
- Identifying the outlier too aggressive.
- Taking a long time to run and computer intensive.

#### Boxplot Outlier Rule

- + Easy to read.
- + Does not depend on estimate of the center of the data.
- May fail if the dataset has many outliers.
- Better suited to distribution that are moderately asymmetric.

