1. **Data Understanding (Understanding the Variables)** 
   1. Education
      * This is a categorical variable with an inherent order (High School < Bachelor’s < Master’s < PhD)
      * Explore the correlation between education levels and salary (see the correlation matrix)
   2. Location
      * No inherent order exists for this categorical variable
      * Explore correlations to assess how location impacts salary (*due to the correlation matrix there is low level of correlation between location and salary*)
   3. Gender
      * Add a dummy variable for gender (e.g., 1 = Female, 0 = Male).

*Due to the correlation matrix, there is low correlation between salary and location / job title / age / gender.*

1. **Analyzing the Salary**

Analyze the distribution of salaries to determine asymmetric. If the distribution is highly asymmetric, consider applying a *logarithmic transformation*. After modeling, apply the anti-log transformation to interpret predictions.

1. **Data Preparation** 
   1. Outlier Analysis
      * Identify and handle outliers using statistical methods.
      * Examine the share of outliers and decide whether to cap, transform, or remove them.
   2. Handling Missing Values
      * Identify the type of missing values (MCAR or MAR). Apply suitable imputation methods or remove rows if the data is not representative.
   3. Data Balance

Check whether the dataset is balanced. If imbalanced, consider oversampling/undersampling techniques.

1. **Modeling and Evaluation**
   1. Model Selection – Linear Regression
      * Use Ridge or Lasso regression to automatically select important variables OR
      * Manually define feature selection criteria based on domain knowledge and correlation analysis.
   2. Evaluation
      * Perform 70%/30% train-test split