1. Check for missing values and handle them appropriately. Use methods like isnull(), fillna(), dropna() to handle missing values.
2. Check for duplicates and handle them appropriately. Use methods like duplicated(), drop_duplicates() to handle duplicates.
3. Check the data types of each column and convert them as necessary. Use methods like dtypes, astype() to convert data types.
4. Check for outliers in the data and handle them appropriately. Use methods like describe(), boxplot(), scatterplot() to identify and handle outliers.
5. Analyze the distribution of the target variable salary and other numerical variables. Use methods like histogram(), density plot(), boxplot() to analyze the distribution of data.
6. Analyze the correlation between the target variable salary and other numerical variables. Use methods like heatmap(), pairplot(), scatterplot() to analyze the correlation between variables.
7. Analyze the relationship between categorical variables and the target variable salary. Use methods like barplot(), boxplot(), violinplot() to analyze the relationship between categorical variables and the target variable.
8. Identify the most important features that affect the target variable. Use methods like feature importance(), correlation analysis(), regression analysis() to identify important features.
9. Perform data pre-processing like scaling, normalization, encoding categorical variables to prepare the data for modeling.
10. Split the data into training and testing datasets.