ALL ABOUT DATA

This report provides an overview of data understanding, manipulation, and preparation for modeling, also known as data preprocessing.

After identifying the business problem, we collaborated with the XYZ Human Resource team to comprehend the meaning of each column in the dataset. Once we understood the features, we utilized Python to analyze the basic descriptive statistics, check for missing values, and examine the dataset's structure. During this process, we created visualizations using Tableau to conduct univariate and bivariate analysis.

Our visualizations revealed some key insights. For example, we observed that top-level employees received three times more pay than mid-level employees, but there were significant pay variations among mid-level employees in different departments.

We addressed missing values by using various techniques, such as filling values from related columns, using the mean, etc. We dropped some columns that were not relevant for modeling or had a low impact. We created derived columns to capture the most information in a single column and removed dependent columns. We also dropped some rows depending on the number of missing values.

We performed manipulations such as converting string columns to integers to prepare the data for modeling. Before proceeding to modeling, we divided the dataset into train, validation, and test sets. We trained the model on the train dataset, evaluated model performance on the validation dataset, and made predictions on the test dataset.

We split the dataset into X and Y, with X representing independent variables and Y representing dependent variables.

After being prepared, we proceeded with the modeling.