**Salary Analysis using Linear Regression**

**Importing Necessary Packages**

In my Jupyter notebook, I began the analysis by importing essential Python packages. This step is crucial as it sets the foundation for data manipulation, visualization, and modeling. The packages used include NumPy, pandas, scikit-learn, and Seaborn.

**Problem Statement**

The main objective of this analysis was to predict employee salaries based on their years of experience. This problem statement is significant for organizations as it enables better workforce planning, salary negotiation, and budget allocation.

**Data Cleaning**

One of the initial steps involved data cleaning. Fortunately, there were no missing or null values in the dataset. This is essential to ensure the quality of the data and prevent any issues during analysis.

**Exploratory Data Analysis (EDA)**

I conducted exploratory data analysis to gain insights into the data's characteristics. Key EDA steps included:

**Regression Plot:** This was used to understand the nature of the relationship between years of experience and salary. It helps in identifying whether the relationship is linear, polynomial, or some other form. Here, in this case, it was Positive Linear Regression and the correlation between the variables is 0.9782.

**Histogram:** A histogram was generated to visualize the distribution of the data. This aids in assessing the normality or skewness of the data. In this case, it was slightly Right-skewed and the skewness value is 0.34.

**Box Plot:** Box plots were used to detect the presence of outliers in the dataset, which can significantly impact model performance. In this case, there is no Outliers.

**Line Plot:** Line plots were created to visualize trends and patterns in the data. In this case, Salary increases with Years of Experience but it's not perfectly Linear.

**Model Building**

Using scikit-learn, I began building the regression model. The "YearsExperience" variable was chosen as the independent variable, while "Salary" served as the dependent variable. The choice of these variables was based on the problem statement and the need to predict salaries accurately.

**Data Splitting**

To evaluate the model's performance, I split the data into training and testing sets. This was done in three different ways:

* 70:30 split.
* 75:25 split.
* 80:20 split.

**Model Selection**

The key objective was to determine which data split resulted in the best model. After evaluating the performance of the models based on R2 score and RMSE values, I found that the 75:25 split produced the most accurate predictions with an R2 score of 94%.