Data Cleaning

Multivariate Imputation

Intuition: We have studied the mean and median methods for handling the missing values present in the dataset. The drawback with these methods is that they only consider a single variable to impute the missing values. But there may be some cases where the missing value in one column is related to the values in other observed columns. We can visualize this with the help of an example.

The sample dataset below contains information about random people who applied for a personal loan.

Name	Age	Experience	Salary (K)	Personal Loan
Roy	25		50	1
Andy	27	3		1
Aksh	29	5	80	0
Hadan	31	7	90	0
Von	33	9	100	1
Tom		11	130	0

Figure: Sample Dataset for Personal Loan

This dataset contains some missing values in various columns. Now, let's try to impute these missing values using univariate methods like mean, median or standard deviation. In this case, we are replacing all the missing values using mean imputation.

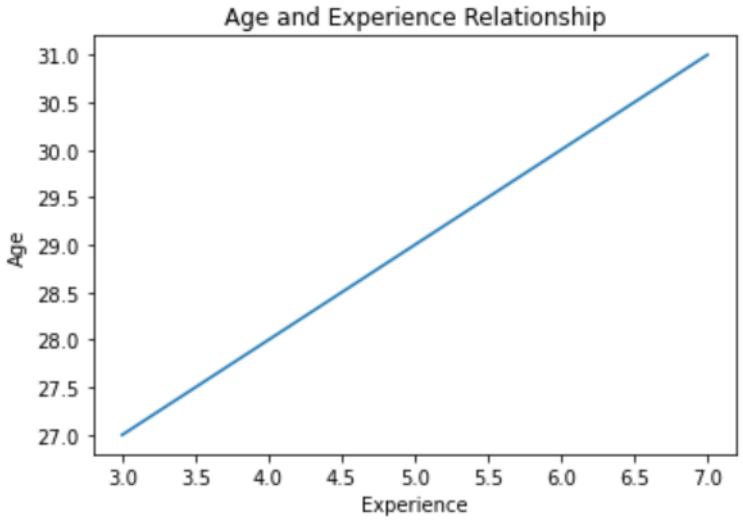
Name	Age	Experience	Salary (K)	Personal Loan
Roy	25	7	50	1
Andy	27	3	90	1
Aksh	29	5	80	0
Hadan	31	7	90	0
Von	33	9	100	1
Tom	29	11	130	0

As you can see that Tom has been imputed with an age of 29 but has an experience of 11 years. If we analyse all other age values related to their experience, it does not match other observations. Andy, who is 27 years old, has experience of 3 years, and Aksh, who is 29 years old, has experience of 5 years. It does not make sense that Tom is 29 years old with 11 years of experience. This is the problem with univariate imputation methods; they do not consider other factors for analysing the values. The multivariate technique solves this issue by factoring in other variables in the data to make better predictions of the missing values.

Using Multivariate Method for Imputation

As the univariate methods do not take other variables into account, we can use multivariate techniques like regression to predict the missing value in one column using other variables. After analyzing the variables, we see that the experience and age columns have a linear relationship.

In [28]:
```
plt.plot(X.reshape(-1, 1), regressor.predict(X.reshape(-1, 1)))
plt.xlabel("Experience")
plt.ylabel("Age")
plt.title("Age and Experience Relationship")
plt.show()
```



We can train a regression model on the experience and age features and then predict the missing age value of “Tom”

In [31]:
```
regressor = LinearRegression()
regressor.fit(X.reshape(-1, 1), y)
print("Predicted Age for Tom is:", round(regressor.predict([[11]])[0]))
```

```
Predicted Age for Tom is: 35
```

The linear regressor trained on “Experience” as X and “Age” as y is predicting 35 as the age of Tom for the experience of 11 years. This seems the correct prediction compared to the dataset's other age and experience values.

Similarly, if we use the “Age” column to predict the missing value of Experience for “Roy”

In [33]:
```
X = df.iloc[1:4, 1].values
y = df.iloc[1:4, 2].values

regressor_experience = LinearRegression()
regressor_experience.fit(X.reshape(-1, 1), y)
print("Predicted Experience for Roy is:", round(regressor_experience.predict([[25]])[0]))
```

```
Predicted Experience for Roy is: 1
```

The linear regressor trained on “Age” as X and “Experience” as y is predicting 1 as the experience of Roy for the age of 25 years. This seems the correct prediction compared to the dataset's other age and experience values.

This seems fun and exciting as we can correctly predict the missing age and experience values using the regression technique. Let's take this a step further and try to predict the missing value in the salary column with the help of the "Age" and "Experience" columns.

```
In [62]:
1 X = df.drop([df.index[1]])[['Age', 'Experience']]
2 y = df.drop([df.index[1]])[['Salary']]
```

```
In [71]:
1 regressor_salary = LinearRegression()
2 regressor_salary.fit(X.values, y)
3
4 predicted_salary = round(regressor_salary.predict([[27, 3]])[0][0])
5
6 print("Predicted Salary for Andy is:", predicted_salary)

Predicted Salary for Andy is: 63
```

As shown in the above code snippet, the salary of Andy is predicted to be 63 K. This seems right compared to the other values. **This is how we can use multivariate techniques to predict the missing values.**

Using MICE

In the next video, we will implement one more multivariate method on the BigMart dataset called the MICE algorithm. This algorithm also does the same multivariate analysis for predicting the missing values. The steps involved in this algorithm are as below:

1. **Remove all the missing values of the selected column.**
2. **Predict the missing values of that column using other columns.**
3. **Similarly, remove all the missing values of column 2.**
4. **Predict the missing values of column 2 using other columns.**
5. **Do it for all the columns for which missing values are present.**