Data Pre-processing

1. Extract Target Variable: Indexing: loc, iloc, name

2. Extract Predictor Variable: Indexing: loc, iloc, names,slicing

3. Train-Test Split

```
From sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(x,y,test_size = 0.20, train_size = 0.80,
random_state = 34)
```

4. How to findout the number of missing values?
   - Isna().sum()
   - Isnull().sum()

5. How to handle missing values?

   a. Fill it Mean(Unskewed), Median(Skewed), Mode(Categorical)

   b. Omission:

   Omit the row if missing values are more in the row (> 80%)
   Omit the column if missing values are more in the column (> 80%)

   c. Backward/Forward fill df.bfill()
   df.ffill()

   d. dropna()

   e. From sklearn import SimpleImputer SimpleImputer(missing_values = np.nan, strategy = ‘mean’)

   f. fillna()

6. Standardizing the data

a. Standard Scalar

```
From sklearn.preprocessing import StandardScalar std = StandardScalar()
std.fit_transform(df['Age','Salary'])
```

b. Min Max Scalar:

```
from sklearn.preprocessing import MinMaxScalar mms = MinMaxScalar()
mms.fit_transform(df['Age','Salary'])
```

7. Encoding the Data

a. Label Encoding:

```
from sklearn.preprocessing import LabelEncoder le = LabelEncoder()
le.fit_transform(df['City'])
```

b. One Hot Encoding

c. Pd.get_dummies()