SALARY DISTRIBUTION

HOW TO DISTINGUISH THE TRAINING DATA AND TEST DATA FROM SALARY DATASET

MACHINE LEARNING

```
import pandas as pd
df=pd.read_csv("/content/salary_data.csv")
df
```

	YearsExperience	Salary
0	1.1	39343
1	1.3	46205
2	1.5	37731
3	2.0	43525
4	2.2	39891
5	2.9	56642
6	3.0	60150
7	3.2	54445
8	3.2	64445
9	3.7	57189
10	3.9	63218
11	4.0	55794
12	4.0	56957
13	4.1	57081
14	4.5	61111
15	4.9	67938
16	5.1	66029
17	5.3	83088
18	5.9	81363
19	6.0	93940
20	6.8	91738
21	7.1	98273
22	7.9	101302
23	8.2	113812
24	8.7	109431
25	9.0	105582
26	9.5	116969
27	9.6	112635
28	10.3	122391
29	10.5	121872

```
df.head()
```

	YearsExperience	Salary
0	1.1	39343
1	1.3	46205
2	1.5	37731
3	2.0	43525
4	2.2	39891

```
df.tail()
```

	YearsExperience	Salary
25	9.0	105582
26	9.5	116969
27	9.6	112635
28	10.3	122391
29	10.5	121872

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30 entries, 0 to 29
Data columns (total 2 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   YearsExperience  30 non-null     float64
 1   Salary           30 non-null     int64  
dtypes: float64(1), int64(1)
memory usage: 608.0 bytes
```

```
df.describe()
```

	YearsExperience	Salary
count	30.000000	30.000000
mean	5.313333	76003.000000
std	2.837888	27414.429785
min	1.100000	37731.000000
25%	3.200000	56720.750000
50%	4.700000	65237.000000
75%	7.700000	100544.750000
max	10.500000	122391.000000

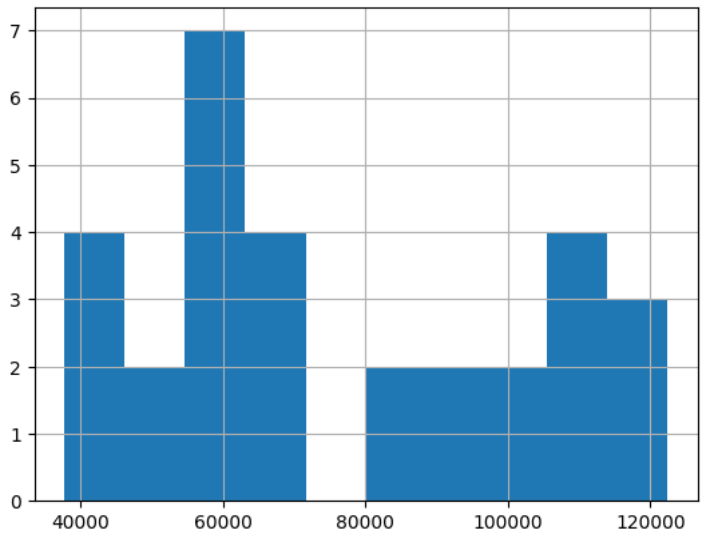
```
df.columns
```

```
Index(['YearsExperience', 'Salary'], dtype='object')
```

```
y=df['Salary']
X=df[['YearsExperience']]
```

```
df['Salary'].hist()
```

```
<Axes: >
```




```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y, train_size=0.7, random_state=2529)
X_train.shape, X_test.shape, y_train.shape, y_test.shape
```

```
((21, 1), (9, 1), (21,), (9,))
```

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(X_train,y_train)
```

```
▾ LinearRegression
LinearRegression()
```

```
model.intercept_
```

```
25249.540132029535
```

```
model.coef_
```

```
array([9441.46935165])
```

```
y_pred = model.predict(X_test)
y_pred
```

```
array([ 37523.45028917,  75289.32769576,  73401.03382543, 102669.58881554,
       110222.76429686,  44132.47883532,  89451.53172323, 114943.49897268,
        39411.7441595 ])
```

```
from sklearn.metrics import mean_absolute_error, mean_absolute_percentage_error, mean_squared_error
```

```
mean_absolute_error(y_test,y_pred)
```

```
5137.291513414335
```

```
mean_absolute_percentage_error(y_test,y_pred)
```

```
0.07066663894000527
```

```
mean_squared_error(y_test,y_pred)
```

```
38750002.34819778
```

✓ 0s completed at 10:40 AM