1. Business Use Case

I want to predict the candidate salary based in the years of experience

Datasets

Data Classification

1. Structured Data -> Excel, CSV, DB,.....
2. Un Structured Data -> Images, Videos, PDF, text files, docs, signal, ...
3. Semi-Structured Data --> xml,html,json,......

Structured Data

- csv - Comma Seperated Values
- tcv - tab seperated values

2. Data Exploration

In [1]:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
```

In [2]:

```
df = pd.read_csv("https://raw.githubusercontent.com/AP-State-Skill-Development-Corporation/
df1 = pd.read_csv("Salary_Data.csv")
```

In [3]:

```
df.head()
```

Out[3]:

	YearsExperience	Salary
0	1.1	39343.0
1	1.3	46205.0
2	1.5	37731.0
3	2.0	43525.0
4	2.2	39891.0

In [4]:

```
df1.tail()
```

Out[4]:

	YearsExperience	Salary
25	9.0	105582.0
26	9.5	116969.0
27	9.6	112635.0
28	10.3	122391.0
29	10.5	121872.0

In [5]:

```
# missing values

df.isnull().sum()
```

Out[5]:

```
YearsExperience    0
Salary             0
dtype: int64
```

In [8]:

```
# Duplicate Values
df.duplicated().sum()
```

Out[8]:

```
0
```

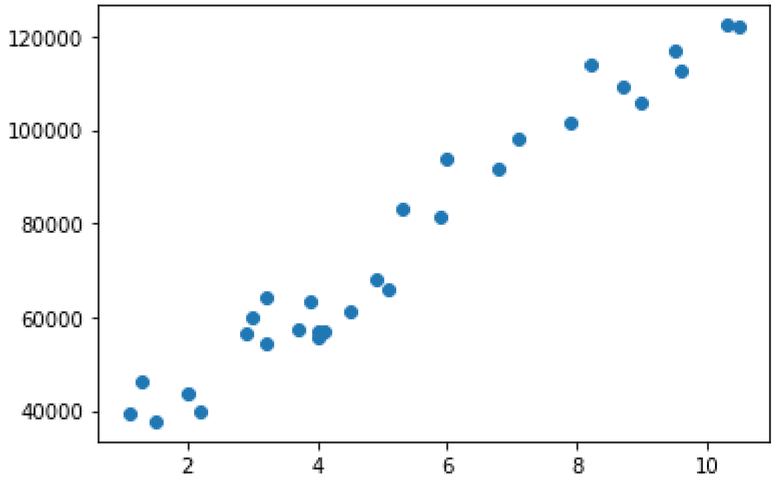
```
1.1 50k per
1.1 30k per month
```

In [9]:

```
plt.scatter(df['YearsExperience'], df['Salary'])
```

Out[9]:

```
<matplotlib.collections.PathCollection at 0x1d8bdeee640>
```



Select Algorithm

Linear Regression

Based on previous plot we can say that `+ve strong Linearly coreleated'

In [10]:

```
## step1: Import Algorithm

from sklearn.linear_model import LinearRegression
```

Build ML Model

In [12]:

```
## Step2 Apply data to the model --> fit the model

model = LinearRegression()

X = df['YearsExperience'].values.reshape(-1, 1)
Y = df['Salary']

model.fit(X,Y)
```

Out[12]:

```
LinearRegression()
```

In [13]:

```
# Step3 predicting our the output

model.predict([[1.1]])
```

Out[13]:

```
array([36187.15875227])
```

In [14]:

```
model.predict([[5.1]])
```

Out[14]:

```
array([73987.00803809])
```

In [15]:

```
y_predict = model.predict(X)
```

Evaluate the model

In [16]:

```
model.score(X, y_predict)
```

Out[16]:

```
1.0
```

In [17]:

```
df.shape
```

Out[17]:

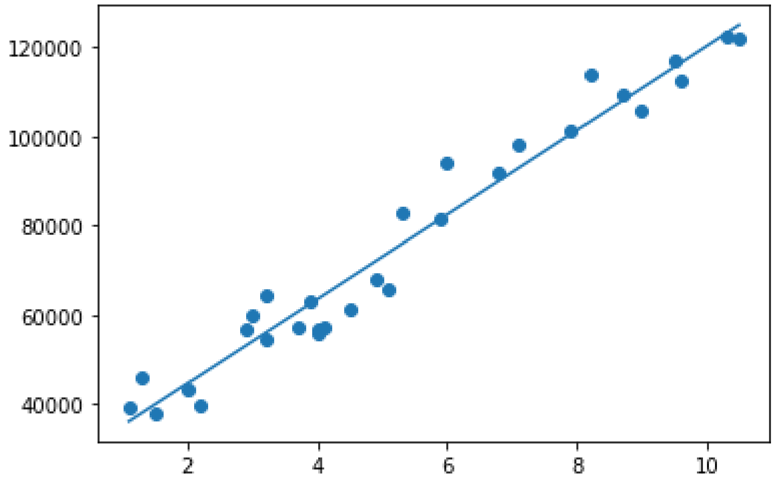
```
(30, 2)
```

In [18]:

```
plt.scatter(df['YearsExperience'], df['Salary'])
plt.plot(df['YearsExperience'], y_predict)
```

Out[18]:

```
[<matplotlib.lines.Line2D at 0x1d8c93b5fd0>]
```



In [19]:

```
model.coef_
```

Out[19]:

```
array([9449.96232146])
```

In [20]:

```
model.intercept_
```

Out[20]:

```
25792.20019866871
```

9449.67 * X + 25792.20 + 0.0