Problem Statement

Linear Regression

Import Libraries

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
a=pd.read_csv("placement.csv")
a
```

Out[2]:

	cgpa	placement_exam_marks	placed
0	7.19	26	1
1	7.46	38	1
2	7.54	40	1
3	6.42	8	1
4	7.23	17	0
...	...	...	...
995	8.87	44	1
996	9.12	65	1
997	4.89	34	0
998	8.62	46	1
999	4.90	10	1

1000 rows × 3 columns

To display top 10 rows

In [3]:
```
c=a.head(15)
c
```

Out[3]:

	cgpa	placement_exam_marks	placed
0	7.19	26	1
1	7.46	38	1
2	7.54	40	1
3	6.42	8	1
4	7.23	17	0

	cgpa	placement_exam_marks	placed
5	7.30	23	1
6	6.69	11	0
7	7.12	39	1
8	6.45	38	0
9	7.75	94	1
10	6.82	16	1
11	6.38	7	1
12	6.58	16	1
13	5.68	26	0
14	7.91	43	0

To find Missing values

In [4]:
```
c.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15 entries, 0 to 14
Data columns (total 3 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   cgpa                 15 non-null     float64
 1   placement_exam_marks  15 non-null     int64  
 2   placed               15 non-null     int64  
dtypes: float64(1), int64(2)
memory usage: 488.0 bytes
```

To display summary of statistics

In [5]:
```
a.describe()
```

Out[5]:

	cgpa	placement_exam_marks	placed
count	1000.000000	1000.000000	1000.000000
mean	6.961240	32.225000	0.489000
std	0.615898	19.130822	0.500129
min	4.890000	0.000000	0.000000
25%	6.550000	17.000000	0.000000
50%	6.960000	28.000000	0.000000
75%	7.370000	44.000000	1.000000
max	9.120000	100.000000	1.000000

To display column heading

In [6]:
```
a.columns
```

Out[6]:
```
Index(['cgpa', 'placement_exam_marks', 'placed'], dtype='object')
```

Pairplot

In [7]:
```
s=a.dropna(axis=1)
s
```

Out[7]:

	cgpa	placement_exam_marks	placed
0	7.19	26	1
1	7.46	38	1
2	7.54	40	1
3	6.42	8	1
4	7.23	17	0
...	...	...	...
995	8.87	44	1
996	9.12	65	1
997	4.89	34	0
998	8.62	46	1
999	4.90	10	1

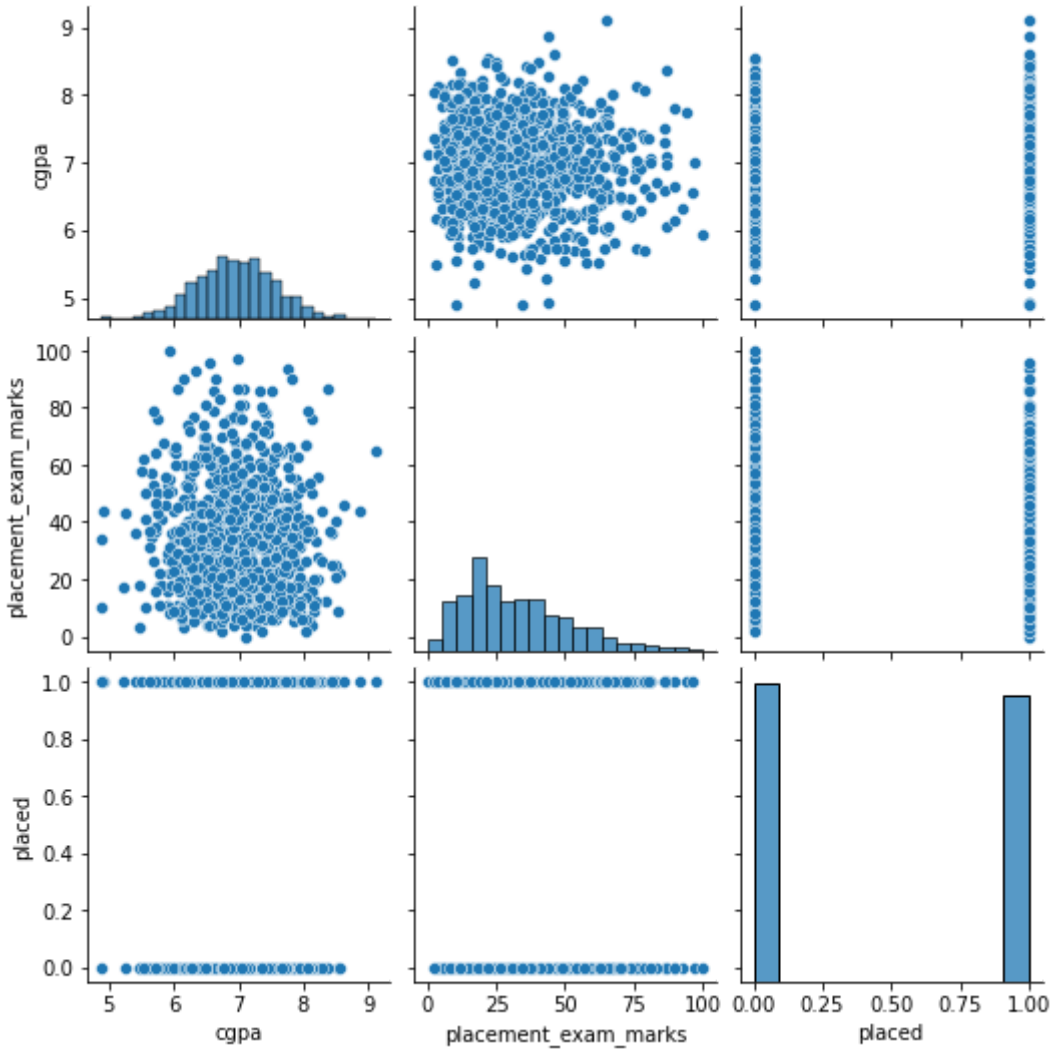
1000 rows × 3 columns

In [8]:
```
s.columns
```

Out[8]:
```
Index(['cgpa', 'placement_exam_marks', 'placed'], dtype='object')
```

In [9]:
```
sns.pairplot(a)
```

Out[9]:
```
<seaborn.axisgrid.PairGrid at 0x1b2432a2640>
```

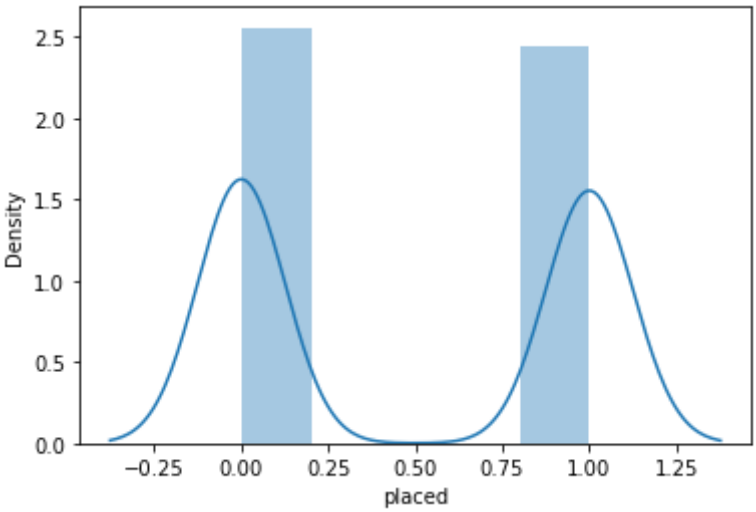


Distribution Plot

In [10]:
```
sns.distplot(a['placed'])
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\distributions.py:2557: FutureWarn
ing: `distplot` is a deprecated function and will be removed in a future version. Pl
ease adapt your code to use either `displot` (a figure-level function with similar f
lexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

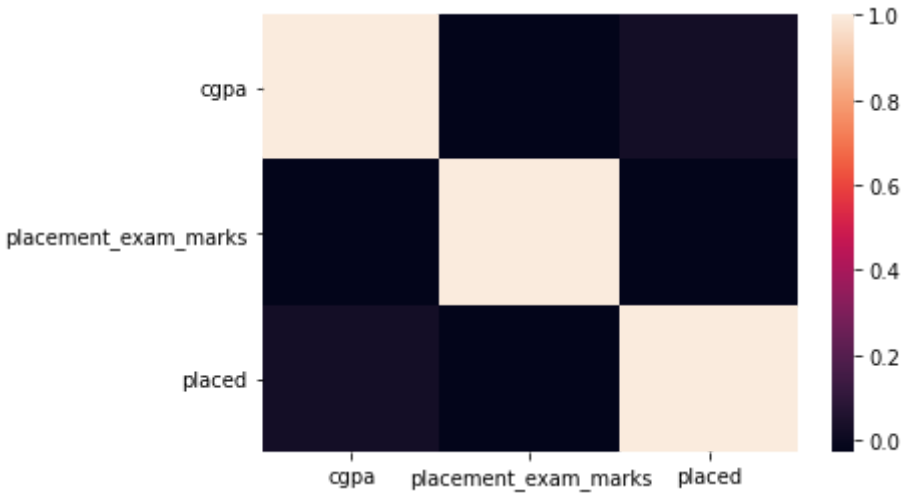
Out[10]: <AxesSubplot:xlabel='placed', ylabel='Density'>



Correlation

In [11]:
```
b=s[['cgpa', 'placement_exam_marks', 'placed']]
sns.heatmap(b.corr())
```

Out[11]: <AxesSubplot:>



Train the model - Model Building

In [12]:
```
g=s[['cgpa', 'placement_exam_marks']]
h=s['placed']
```

To split dataset into training end test

In [13]:
```
from sklearn.model_selection import train_test_split
g_train,g_test,h_train,h_test=train_test_split(g,h,test_size=0.6)
```

To run the model

In [14]:
```
from sklearn.linear_model import LinearRegression
```

In [15]:
```
lr=LinearRegression()
lr.fit(g_train,h_train)
```

Out[15]: LinearRegression()

In [16]:
```
print(lr.intercept_)
```

```
0.18431968886105055
```

Coeffecient

In [17]:
```
coeff=pd.DataFrame(lr.coef_,g.columns,columns=['Co-effecient'])
coeff
```

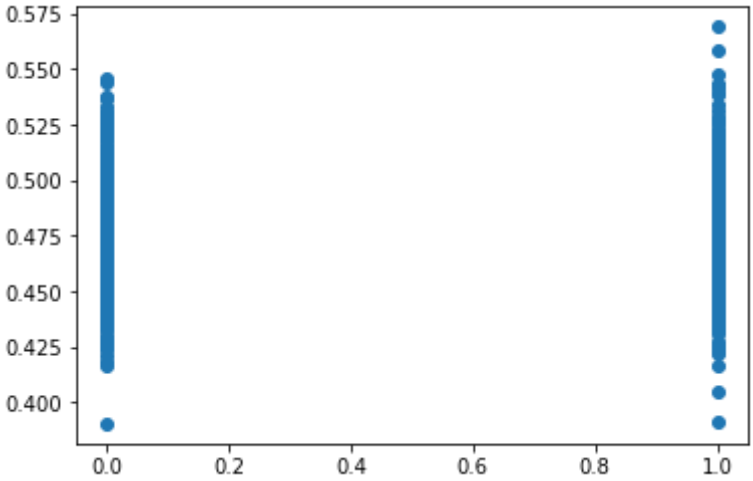
Out[17]:

	Co-effecient
cgpa	0.042196
placement_exam_marks	-0.000002

Best Fit line

In [18]:
```
prediction=lr.predict(g_test)
plt.scatter(h_test,prediction)
```

Out[18]:
```
<matplotlib.collections.PathCollection at 0x1b245a050d0>
```



To find score

In [19]:
```
print(lr.score(g_test,h_test))
```

```
-0.0028892732250840325
```

In [ ]: