7. Test The analysis of varaince between etest_p and mbs_p at significance level 5% (Make decisions using Hypothesis Testing).

In [1]:
```
import pandas as pd
```

In [2]:
```
dataset=pd.read_csv("Preplacementdata.csv")
dataset
```

Out[2]:

	sl_no	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary	gender	ssc_b	hsc_b	hsc_s	degree_t	workex	special
0	1.0	67.00	91.00	58.00	55.0	58.80	270000.0	M	Others	Others	Commerce	Sci&Tech	No	M
1	2.0	79.33	78.33	77.48	86.5	66.28	200000.0	M	Central	Others	Science	Sci&Tech	Yes	M
2	3.0	65.00	68.00	64.00	75.0	57.80	250000.0	M	Central	Central	Arts	Comm&Mgmt	No	M
3	4.0	56.00	52.00	52.00	66.0	59.43	265000.0	M	Central	Central	Science	Sci&Tech	No	M
4	5.0	85.80	73.60	73.30	96.8	55.50	425000.0	M	Central	Central	Commerce	Comm&Mgmt	No	M
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
210	211.0	80.60	82.00	77.60	91.0	74.49	400000.0	M	Others	Others	Commerce	Comm&Mgmt	No	M
211	212.0	58.00	60.00	72.00	74.0	53.62	275000.0	M	Others	Others	Science	Sci&Tech	No	M
212	213.0	67.00	67.00	73.00	59.0	69.72	295000.0	M	Others	Others	Commerce	Comm&Mgmt	Yes	M
213	214.0	74.00	66.00	58.00	70.0	60.23	204000.0	F	Others	Others	Commerce	Comm&Mgmt	No	M
214	215.0	62.00	58.00	53.00	89.0	60.22	265000.0	M	Central	Others	Science	Comm&Mgmt	No	M

215 rows × 15 columns

In [3]:
```
import scipy.stats as stats
```

In [5]:
```
stats.f_oneway(dataset['etest_p'],dataset['mba_p'])
```

Out[5]:
```
F_onewayResult(statistic=98.64487057324706, pvalue=4.672547689133573e-21)
```

null hypothesis H0

There is no differnce between pass mark of etest and mba

Alternate hypothesis H1

There is differnce between pass mark of etest and mba

The calculated p_value is less than 0.05 , we reject the null hypothesis, So the we conclude there is differences between pass marks of etest and mba.

8. Test the similarity between the degree_t(sci & tech) and specialization level of 5%.(make decisions using Hypothesis Testing).

In [20]:
```
from scipy.stats import ttest_ind
degree_tST= dataset[dataset ['degree_t']=="Sci&Tech"]["salary"]
specialisation= dataset[dataset['specialisation']=="MKt&HR"]["salary"]
ttest_ind(degree_tST,specialisation)
```

Out[20]:
```
TtestResult(statistic=nan, pvalue=nan, df=nan)
```

Null Hypothesis (H_0): P_value is less than 0.05

Ther is no significance the between the degree_t(Sci&tech) and specialisation(Mkt&HR) with respect to salary

Alternate Hypothesis (H_a):

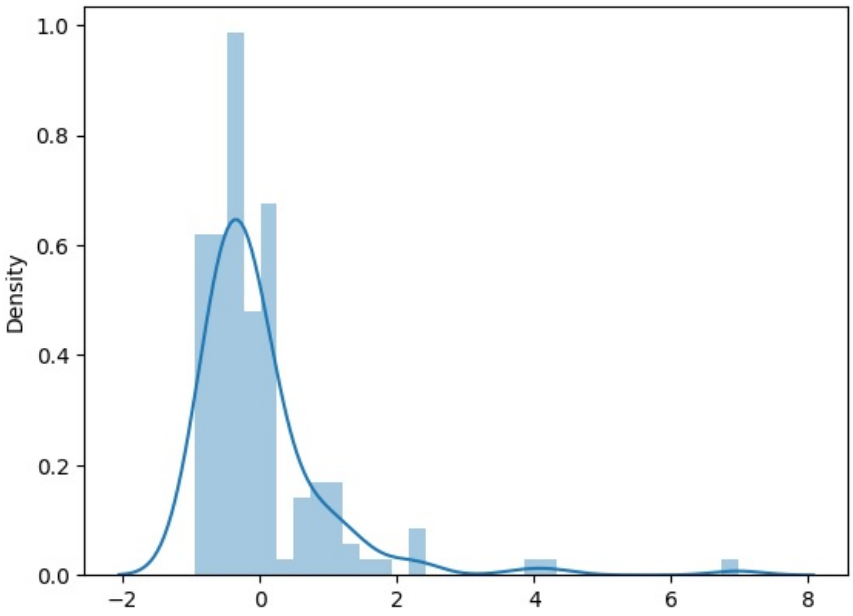
Ther is no significance between the degree_t(Sci&tech) and specialisation (Mkt&HR) with respect to salary

9. Convert the normal distribution to standard normal distribution for the salary column.

In [28]:
```
def stdNBgraph(dataset):
    import seaborn as sns
    mean=dataset.mean()
    std=dataset.std()
    values=[i for i in dataset]
    z_score=[((j-mean)/std) for j in values]
    sns.distplot(z_score,kde=True)
    sum(z_score)/len(z_score)
```

In [29]:
```
stdNBgraph(dataset["salary"])
```

```
C:\Users\SowmiGanesh\AppData\Local\Temp\ipykernel_10228\1411587287.py:7: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(z_score,kde=True)
```



10. What is the probability Density Function of the salary range from 700000 to 900000?

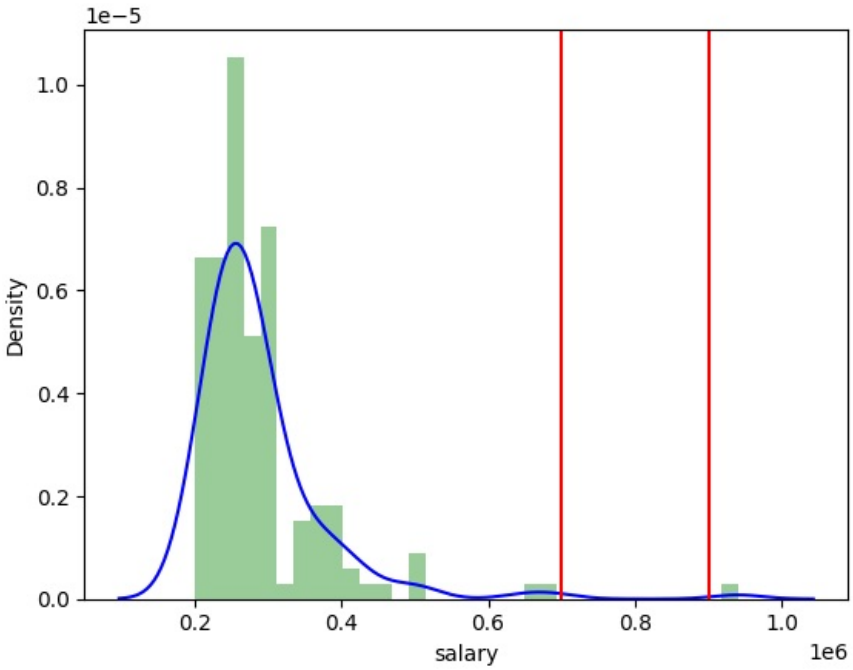
In [30]:
```
def get_pdf_probability(dataset,startrange,endrange):
    from matplotlib import pyplot
    from scipy.stats import norm
    import seaborn as sns
    ax = sns.distplot(dataset,kde=True,kde_kws={'color':'blue'},color='Green')
    pyplot.axvline(startrange,color='Red')
    pyplot.axvline(endrange,color='Red')
    # generate a sample
    sample = dataset
    # calculate parameters
    sample_mean =sample.mean()
    sample_std = sample.std()
    print('Mean=%.3f, Standard Deviation=%.3f' % (sample_mean, sample_std))
    # define the distribution inbulit function
    dist = norm(sample_mean, sample_std)

    # sample probabilities for a range of outcomes ( for loop to list single line it convert to a list)
    values = [value for value in range(startrange, endrange)]
```

```
    probabilities = [dist.pdf(value) for value in values]
    prob=sum(probabilities)
    print("The area between range({},{}):{}".format(startrange,endrange,sum(probabilities)))
    return prob
```

In [31]:
```
get_pdf_probability(dataset["salary"], 700000, 900000)
```

```
C:\Users\SowmiGanesh\AppData\Local\Temp\ipykernel_10228\3298601999.py:5: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  ax = sns.distplot(dataset,kde=True,kde_kws={'color':'blue'},color='Green')
Mean=288655.405, Standard Deviation=93457.452
The area between range(700000,900000):5.377578376230696e-06
```

Out[31]:
```
5.377578376230696e-06
```



11. Test the similarity between the degree_t(sci& tech) with respect to etest_p and mba_p at significance level of 5%. (make decisions using Hypothesis testing).

Null hypothesis(H0):

there is no significance difference between the degree_t(Sci&tech) with respect to etest_p and mba_p.

Alternative hypothesis(Ha):

there is a significant difference between the degree_t(Sci&tech) with respect to etest_p and mba_p.

Test statistics:

In [36]:
```
from scipy.stats import ttest_rel
degree_tet=dataset[ dataset['degree_t']=="Sci&tech"]["etest_p"]
degree_tmt=dataset[ dataset['degree_t']=="Sci&tech"]["mba_p"]
ttest_rel(degree_tet,degree_tmt)
```

Out[36]:
```
TtestResult(statistic=nan, pvalue=nan, df=nan)
```

In [ ]:
```

```

In [ ]:

In [ ]: