Data Science assignment

1) **Replace the NaN values with the correct value. And justify why you have chosen the same.**

To Find NaN values in Python: dataset.isnull().sum()

Handling NaN Values - Four common ways to handle NaN
1. Replace with mean, median, or mode
2. Replace with 0 or a constant value
3. Drop rows with NaN values: Simple, but can result in data loss.
4. Use machine learning to predict missing values

Justification: In the case of salary, using methods 1, 3, or 4 would lead to incorrect implications (e.g., unplaced candidates having a valid salary). Thus, replacing NaN with 0 (method 2) is the best choice, as it accurately reflects that unplaced candidates have no salary.

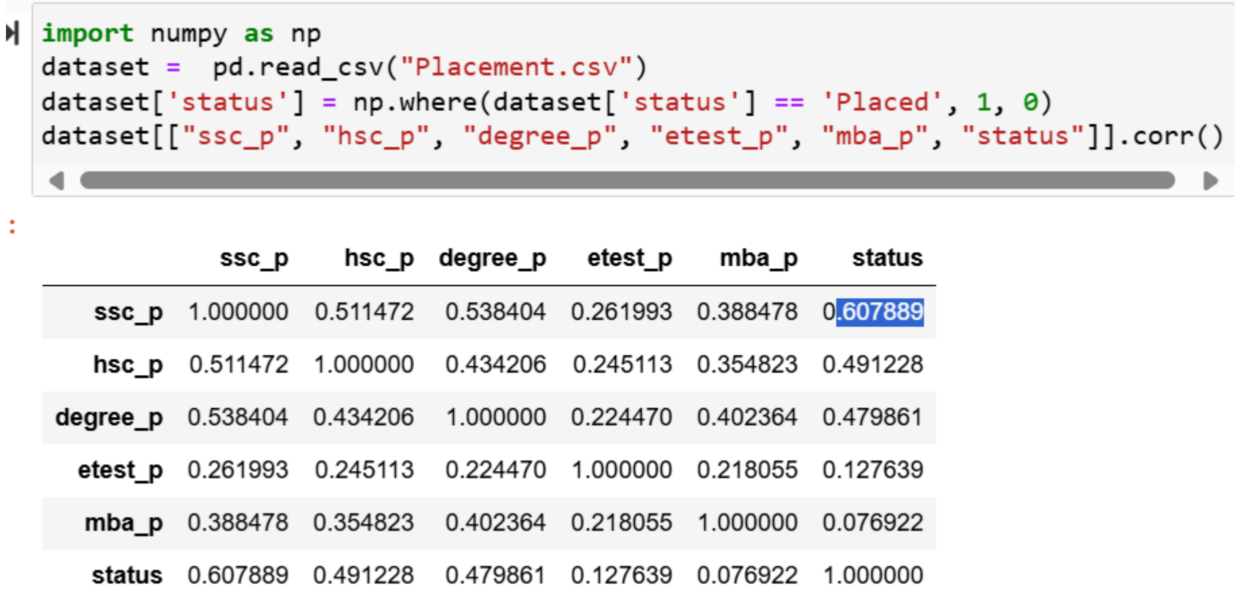
2) **How many of them are not placed?**

There are 67 candidates who are not placed, as determined by analyzing the `'status'` column of the dataset.

# Frequency of ‘Not placed’ candidates
notPlacedCount =dataset['status'].value_counts().get('Not Placed', 0)

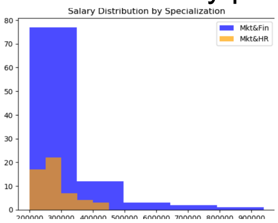
3) **Find the reason for non placement from the dataset?**

`ssc_p` stands out as the most critical factor. Candidates with lower secondary school scores are significantly less likely to be placed.



```
import numpy as np
dataset =  pd.read_csv("Placement.csv")
dataset['status'] = np.where(dataset['status'] == 'Placed', 1, 0)
dataset[["ssc_p", "hsc_p", "degree_p", "etest_p", "mba_p", "status"]].corr()
```

	ssc_p	hsc_p	degree_p	etest_p	mba_p	status
ssc_p	1.000000	0.511472	0.538404	0.261993	0.388478	0.607889
hsc_p	0.511472	1.000000	0.434206	0.245113	0.354823	0.491228
degree_p	0.538404	0.434206	1.000000	0.224470	0.402364	0.479861
etest_p	0.261993	0.245113	0.224470	1.000000	0.218055	0.127639
mba_p	0.388478	0.354823	0.402364	0.218055	1.000000	0.076922
status	0.607889	0.491228	0.479861	0.127639	0.076922	1.000000

4)	**What kind of relation between salary and mba_p**
	They are positive correlated with a weaker degree. When the mba_p score increases there is a very small tendency for increase in salary. Correlation coefficient between salary and mba_p = 0.175 dataset[['salary','mba_p']].corr()
5)	**Which specialization is getting a minimum salary?**
	'Mkt&Fin' and 'Mkt&HR' both get the lowest salary of 200000. There are more no. of candidates getting the lowest salary compared to Mkt&HR candidates This is found by plotting the histogram 
6)	**How many of them are getting above 500,000 salary?**
	3 candidates get a salary greater than 500,000. This is found using Frequency - value_counts() (dataset['salary']>-1).value_counts()
7)	**Test the Analysis of Variance between etest_p and mba_p at significance level 5%.(Make decision using Hypothesis Testing)**
	pvalue=4.672547689133573e-21 P value is very much less than 5% Reject the Null Hypothesis (H0): There is sufficient evidence based on pvalue to conclude that there is a significant difference between the means of `etest_p` and `mba_p`. Used One-way classification - stats.f_oneway(dataset['etest_p'],dataset['mba_p'])
8)	**Test the similarity between the degree_t(Sci&Tech) and specialisa tion(Mkt&HR) with respect to salary at significance level of 5%.(Make decision using Hypothesis Testing)**
	○ Degree type (Sci&Tech) does not significantly affect salary. ○ Specialisation (Mkt&HR) significantly affects salary. ○ The interaction between degree type (Sci&Tech) and specialisation (Mkt&HR) does not significantly affect salary. Hypothesis Testing to Compare Degree Type (Sci&Tech) and Specialisation (Mkt&HR) with Respect to Salary: We are testing whether degree type (Sci&Tech) and specialisation (Mkt&HR),

along with their interaction, have a significant effect on salary, using Two-Way ANOVA at a 5% significance level.

Step 1: Hypotheses

1. Null Hypothesis (H0H_0):
   - There is no significant effect of:
     - `degree_t` (Sci&Tech) on salary,
     - `specialisation` (Mkt&HR) on salary,
     - and the interaction between `degree_t` and `specialisation` on salary.
2. Alternative Hypothesis (HaH_a):
   - At least one of the factors (`degree_t`, `specialisation`, or their interaction) has a significant effect on salary.

Step 2: Python Code to Perform Two-Way ANOVA

```
 # Fit the two-way ANOVA model

model = ols('salary ~ C(degree_t) + C(specialisation) + C(degree_t):C(specialisation)',
data=dataset).fit()

anova_table = sm.stats.anova_lm(model, typ=2)
```

Step 3: Decision Rule

1. Significance Level (α\alpha): 5% (0.05).
2. Reject H0H_0 for any factor if the p-value < 0.05.
3. Fail to reject H0H_0 for any factor if the p-value ≥ 0.05.

Step 4: Output (ANOVA Table)

	P-Value
C(degree_t)	0.092033
C(specialisation)	0.059518
C(degree_t):C(specialisation)	0.929838

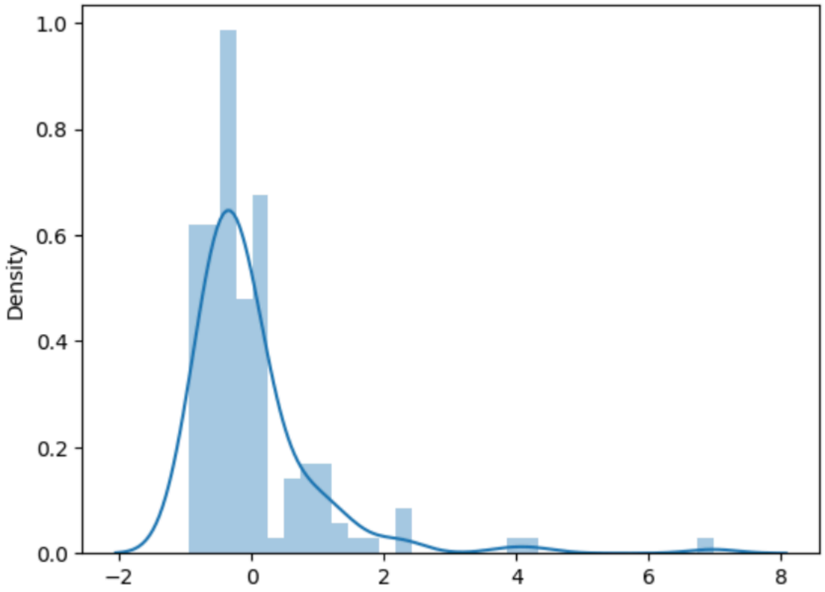
Step 5: Decision and Conclusion

1. Degree Type (`C(degree_t)`):
   - P-value = 0.09 (greater than 0.05).
   - Decision: Fail to reject H0H_0. Conclusion: Degree type (Sci&Tech) does not significantly affect salary.
2. Specialisation (`C(specialisation)`):
   - P-value = 0.059 (greater than 0.05).
   - Decision: Fail to Reject H0H_0. Conclusion: Specialisation (Mkt&HR) does not significantly affect salary.

3. Interaction (`C(degree_t):C(specialisation)`):
   - P-value = 0.929 (greater than 0.05).
   - Decision: Fail to reject H0H_0. Conclusion: The interaction between degree type (Sci&Tech) and specialisation (Mkt&HR) does not significantly affect salary.

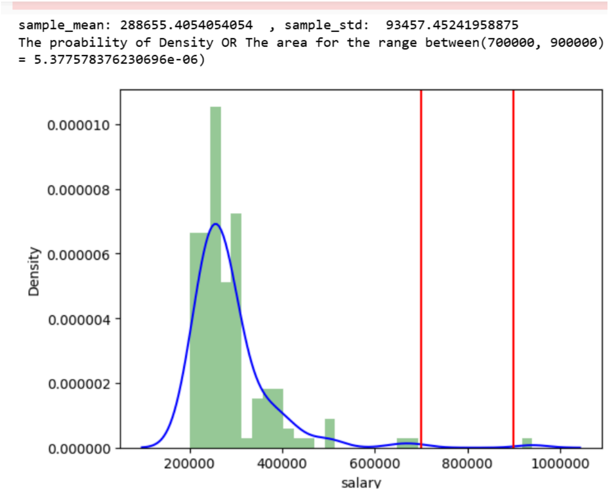
9) **Convert the normal distribution to standard normal distribution for salary column**

Using Z-score with the column's mean and std deviation values we can convert to a standard normal deviation.



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

About 5 - 6 candidates are in the salary range 700000 to 900000



```
sample_mean: 288655.4054054054  , sample_std:  93457.45241958875
The proability of Density OR The area for the range between(700000, 900000)
= 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 decision using Hypothesis Testing)**

There is a significant difference between etest_p and mba_p for candidates with degree_t =

Sci&Tech.
1 group value,
Paired T-test Dependent Sample

```
sciTech_etest = dataset[dataset['degree_t']=='Sci&Tech']['etest_p']
sciTech_mba = dataset[dataset['degree_t']=='Sci&Tech']['mba_p']
```

1. Hypothesis Statements
   - Null Hypothesis (H0H_0): There is no significant difference between the means of etest_p and mba_p for candidates with degree_t = Sci&Tech.
   - Alternative Hypothesis (HaH_a): There is a significant difference between the means of etest_p and mba_p for candidates with degree_t = Sci&Tech.
2. Significance Level: The significance level (α\alpha) is 5% (0.05).
3. Paired T-Test: The paired t-test is used as the two datasets (etest_p and mba_p) are related (collected from the same candidates).

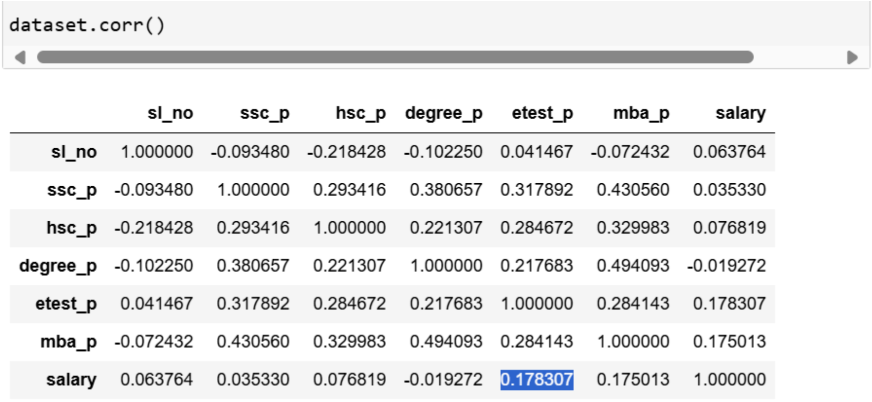
   ```
   sciTech_etest = dataset[dataset['degree_t'] == 'Sci&Tech']['etest_p']
   sciTech_mba = dataset[dataset['degree_t'] == 'Sci&Tech']['mba_p']

   # Perform the paired t-test
   t_stat, p_value = ttest_rel(sciTech_etest, sciTech_mba)
   ```

4. Test Results: T-Statistic: 4.915474373730152. P-Value: 1.5494422054952274e-05
5. Decision Rule
   - If p-value<0.05p, reject the null hypothesis (H0).
   - If p-value≥0.05p, fail to reject the null hypothesis (H0).
6. Conclusion
   A. The p-value = 0.0000155 is much smaller than 0.05.
   B. Thus, we reject the null hypothesis (H0H_0).
   C. There is sufficient evidence to conclude that there is a significant difference between etest_p and mba_p for candidates with degree_t = Sci&Tech.

12) Which parameter is highly correlated with salary?

There is not any highly correlated value with Salary

The positive correlated value for Salary is 17.8% with etest but this is weak degree of positive correlation



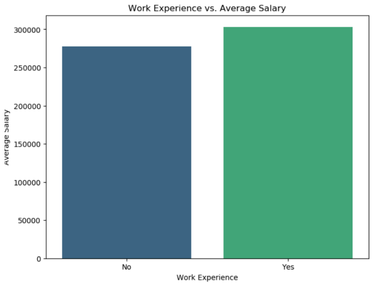
```
dataset.corr()
```

	sl_no	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
sl_no	1.000000	-0.093480	-0.218428	-0.102250	0.041467	-0.072432	0.063764
ssc_p	-0.093480	1.000000	0.293416	0.380657	0.317892	0.430560	0.035330
hsc_p	-0.218428	0.293416	1.000000	0.221307	0.284672	0.329983	0.076819
degree_p	-0.102250	0.380657	0.221307	1.000000	0.217683	0.494093	-0.019272
etest_p	0.041467	0.317892	0.284672	0.217683	1.000000	0.284143	0.178307
mba_p	-0.072432	0.430560	0.329983	0.494093	0.284143	1.000000	0.175013
salary	0.063764	0.035330	0.076819	-0.019272	0.178307	0.175013	1.000000

13) **Plot any useful graph and explain it**

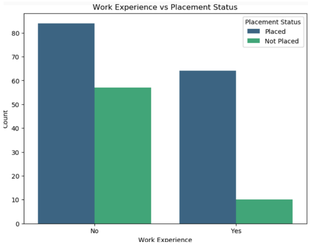
13 A) Does candidates with work experience get higher salary

Ans: Yes



13 B) Do candidates with work experience relate to placed status.

Ans: No. It is not necessary that the candidate with experience get placed sooner.



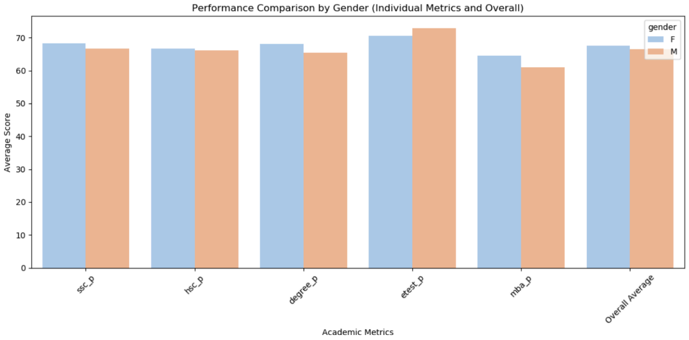
13) C) Which Gender has the overall highest academic performance - ssc_p, hsc_p, degree_p, etest_p, mba_

Ans: Female

Overall Average:

Female: 67.656395%

Male: 66.451007%


Performance Comparison by Gender (Individual Metrics and Overall)
gender
F
M
Average Score
ssc_p
hsc_p
degree_p
etest_p
mba_p
Overall Average
Academic Metrics