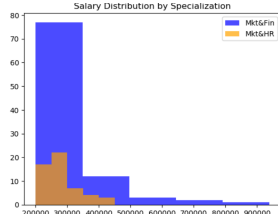


Data Science assignment

1)	<p>Replace the NaN values with the correct value. And justify why you have chosen the same</p> <p>To Find NaN values in Python: dataset.isnull().sum()</p> <p>Handling NaN Values - Four common ways to handle NaN</p> <ol style="list-style-type: none">1. Replace with mean, median, or mode2. Replace with 0 or a constant value3. Drop rows with NaN values: Simple, but can result in data loss.4. Use machine learning to predict missing values <p>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.</p>																																																	
2)	<p>How many of them are not placed?</p> <p>There are 67 candidates who are not placed, as determined by analyzing the 'status' column of the dataset.</p> <p># Frequency of 'Not placed' candidates notPlacedCount =dataset['status'].value_counts().get('Not Placed', 0)</p>																																																	
3)	<p>Find the reason for non placement from the dataset?</p> <p>ssc_p stands out as the most critical factor. Candidates with lower secondary school scores are significantly less likely to be placed.</p> <pre>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()</pre> <p>:</p> <table><tr><th></th><th>ssc_p</th><th>hsc_p</th><th>degree_p</th><th>etest_p</th><th>mba_p</th><th>status</th></tr><tr><th>ssc_p</th><td>1.000000</td><td>0.511472</td><td>0.538404</td><td>0.261993</td><td>0.388478</td><td>0.607889</td></tr><tr><th>hsc_p</th><td>0.511472</td><td>1.000000</td><td>0.434206</td><td>0.245113</td><td>0.354823</td><td>0.491228</td></tr><tr><th>degree_p</th><td>0.538404</td><td>0.434206</td><td>1.000000</td><td>0.224470</td><td>0.402364</td><td>0.479861</td></tr><tr><th>etest_p</th><td>0.261993</td><td>0.245113</td><td>0.224470</td><td>1.000000</td><td>0.218055</td><td>0.127639</td></tr><tr><th>mba_p</th><td>0.388478</td><td>0.354823</td><td>0.402364</td><td>0.218055</td><td>1.000000</td><td>0.076922</td></tr><tr><th>status</th><td>0.607889</td><td>0.491228</td><td>0.479861</td><td>0.127639</td><td>0.076922</td><td>1.000000</td></tr></table>		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
	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)	<p>What kind of relation between salary and mba_p</p> <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</p> <p><code>dataset[['salary','mba_p']].corr()</code></p>
5)	<p>Which specialization is getting a minimum salary?</p> <p>'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</p> <p>This is found by plotting the histogram</p> 
6)	<p>How many of them are getting above 500,000 salary?</p> <p>3 candidates get a salary greater than 500,000.</p> <p>This is found using Frequency - value_counts() <code>(dataset['salary']>500000).value_counts()</code></p>
7)	<p>Test the Analysis of Variance between etest_p and mba_p at significance level 5%.(Make decision using Hypothesis Testing)</p> <p>pvalue=4.672547689133573e-21</p> <p>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.</p> <p>Used One-way classification - <code>stats.f_oneway(dataset['etest_p'],dataset['mba_p'])</code></p>
8)	<p>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)</p> <ul style="list-style-type: none"> ○ 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. <p>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</p>

(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 (H_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 (H_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

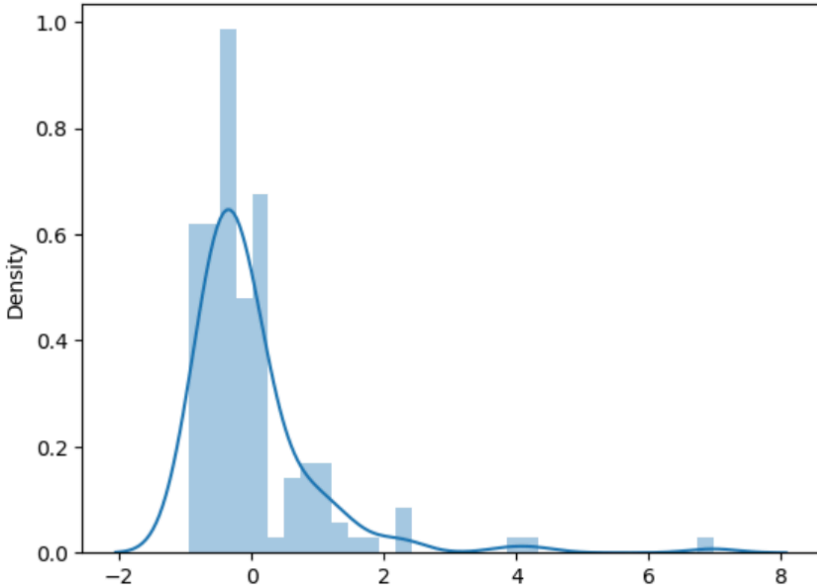
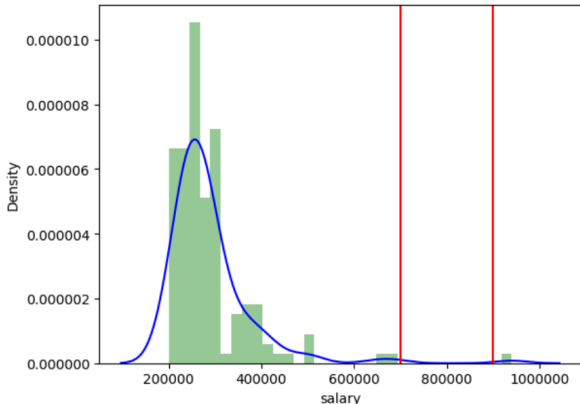
1. Significance Level (α): 5% (0.05).
2. Reject H_0 for any factor if the p-value < 0.05.
3. Fail to reject H_0 for any factor if the p-value \geq 0.05.

Step 4: Output (ANOVA Table)

	P-Value
C(<code>degree_t</code>)	0.092033
C(<code>specialisation</code>)	0.059518
C(<code>degree_t</code>):C(<code>specialisation</code>)	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 H_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 H_0 . Conclusion: Specialisation (Mkt&HR) does not significantly affect salary.

	<p>3. Interaction (C(degree_t):C(specialisation)):</p> <ul style="list-style-type: none"> ○ P-value = 0.929 (greater than 0.05). ○ Decision: Fail to reject H₀. Conclusion: The interaction between degree type (Sci&Tech) and specialisation (Mkt&HR) does not significantly affect salary.
9)	<p>Convert the normal distribution to standard normal distribution for salary column</p> <p>Using Z-score with the column's mean and std deviation values we can convert to a standard normal deviation.</p> 
10)	<p>What is the probability Density Function of the salary range from 700000 to 900000?</p> <p>About 5 - 6 candidates are in the salary range 700000 to 900000</p> <p>sample_mean: 288655.4054054054 , sample_std: 93457.45241958875 The probability of Density OR The area for the range between(700000, 900000) = 5.377578376230696e-06)</p> 
11)	<p>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)</p> <p>There is a significant difference between etest_p and mba_p for candidates with degree_t =</p>

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 (H_0): There is no significant difference between the means of etest_p and mba_p for candidates with degree_t = Sci&Tech.
- Alternative Hypothesis (H_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) 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 $p\text{-value} < 0.05$, reject the null hypothesis (H_0).
- If $p\text{-value} \geq 0.05$, fail to reject the null hypothesis (H_0).

6. Conclusion

A. The p-value = 0.0000155 is much smaller than 0.05.

B. Thus, we reject the null hypothesis (H_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

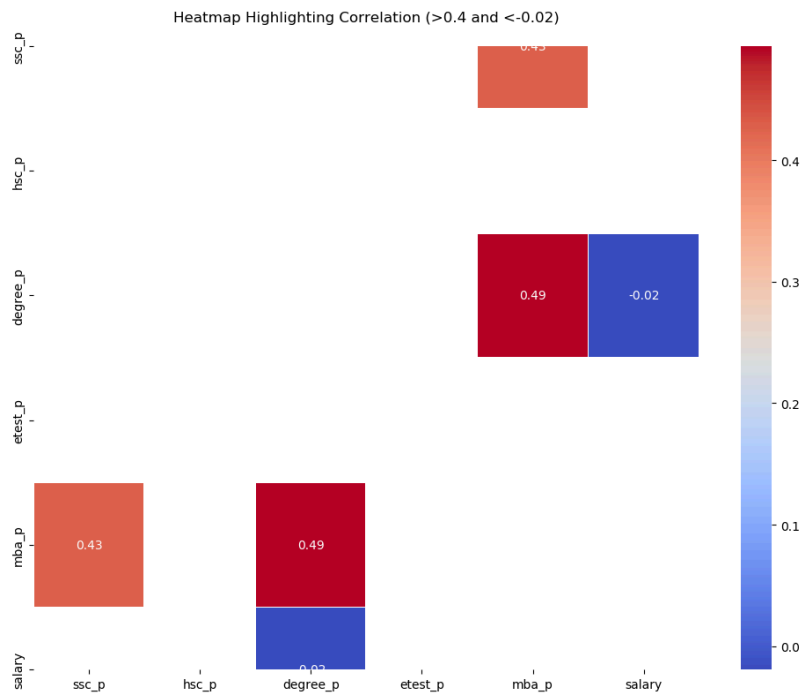
The positive correlated value for Salary is 17.8% with etest but this is weak degree of positive correlation

```
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



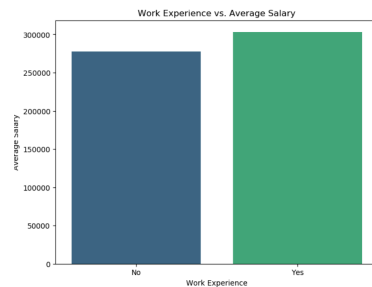
ssc_p and **mba_p**: Medium positive correlation (0.49). Higher secondary school scores lead to higher MBA scores.

degree_p and **salary**: Very weak negative correlation (-0.02). No meaningful link; salary isn't influenced by degree performance.

13 A)

Does candidates with work experience get higher salary

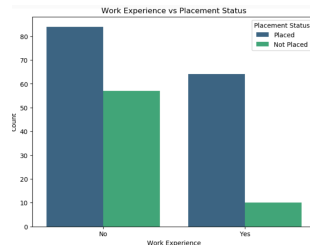
Yes



13 B)

Do candidates with work experience relate to placed status.

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_

Female

Overall Average:

Female: 67.656395%

Male: 66.451007%

