Data Analysis Assignment

Justification or answer summary needed for every question

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

Replaced the Nan values in salary column with 0 because those are not placed in any company so salary value is 0.

- If NaN represents missing information, keeping Nan could make data analysis difficult.
- Many statistical and machine learning algorithms do not handle NaN values well.
- Replacing it with 0 ensures uniformity across the dataset.
- For example, if the dataset includes fresh graduates who don’t have a salary yet, using 0 accurately reflects their situation.
- `dataset["salary"] = dataset["salary"].fillna(0)`

1.3 2) How many of them are not placed?

```
print((dataset['status']).value_counts())
```

output:

```
status
Placed        148
Not Placed     67
```

1.4 3) Find the reason for non-placement from the dataset?

Factors that could affect placement status:

- Academic Performance: Low percentage in MBA or entrance tests.
- Skill Gaps: Lack of technical or soft skills.

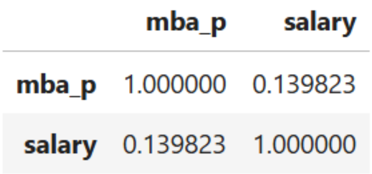
- Interview Performance: Poor aptitude, technical, or HR round performance.
- Extracurricular Activities: Lack of participation in workshops, hackathons, etc.
- Company Preferences: Some companies may have higher cut-offs for marks.

1.5 4) What kind of relation between salary and mba_p?

A Positive Correlation means that as one variable increases, the other variable also increases**.** It indicates a direct relationship between two factors.

```
dataset[['mba_p','salary']].corr()
```

output:



	mba_p	salary
mba_p	1.000000	0.139823
salary	0.139823	1.000000

- correlation difference between mba_p and salary is low positive correlation as it is between 0 to 1.

1.6 5) Which specialization is getting minimum salary?

```
Specialisation with Minimum Non-Zero Salary:
specialisation     Mkt&Fin
salary              200000.0
```

1.7 6) How many of them getting above 500000 salary?

```
print ((dataset["salary"]>=500000).value_counts())
```

```
output:   salary-     False    209
                      True      6
```

1.8 7) Test the Analysis of Variance between etest_p and mba_p at signifance level 5%.(Make decision using Hypothesis Testing)

ANOVA is a statistical method used to compare the means of multiple groups to determine if there is a significant difference between them. It helps in understanding whether variations in the data are due to real differences or just random chance.

Output: statistic=98.64487057324706, pvalue=4.672547689133573e-21

- According to condition p value is 4.67 > 0.05 so accept Null Hypothesis and reject Alternate Hypothesis
- There is no significant difference between etest_p and mba_p

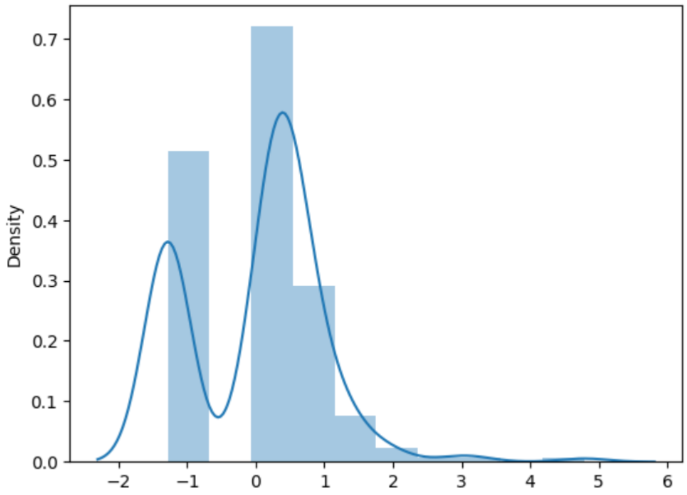
1.9 8)Test the similarity between the degree_t (Sci&Tech) and specialisation (Mkt&HR) with respect to salary at significance level of 5%.(Make decision using Hypothesis Testing)

It is found using Unpaired T-Test – Independent sample

Output: statistic=2.692041243555374, pvalue=0.007897969943471179, df=152.0

- According to condition p value is 0.007 < 0.05 so accept Alternate Hypothesis and reject Null Hypothesis
- There is significant difference between degree_p (Sci&Tech) and specialisation (Mkt&HR) with respect to salary

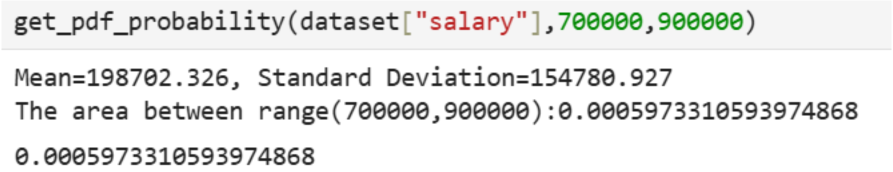
1.10 9) Convert the normal distribution to standard normal distribution for salary column



- The given distribution plot shows the transformation of normal distribution into standard normal distribution (also called Z-score normalization or standardization).
- A standard normal distribution is a normal distribution that says data values are converted into Z-scores, which indicate how many standard deviations a value is away from the mean.
- The x-axis values are now Z-scores (ranging from -2 to +5) instead of actual salary values.
- A peak is seen around Z = 0, meaning most values are close to the mean.
- Extreme values (outliers) appear on the right side (positive Z-scores)
- Allows Comparisons across Different Scales → Useful when combining different datasets.
- Required for Machine Learning Models → Algorithms like logistic regression, SVM, and k-means clustering work better with standardized data.
- Identifies Outliers Easily → Z-scores greater than +3 or less than -3 indicate outliers.

- Essential for Probability Calculations → Helps in computing probabilities using the standard normal table.
- Standardization is a powerful technique to normalize data while preserving its distribution.
- It is widely used in statistics, machine learning, and data analysis.

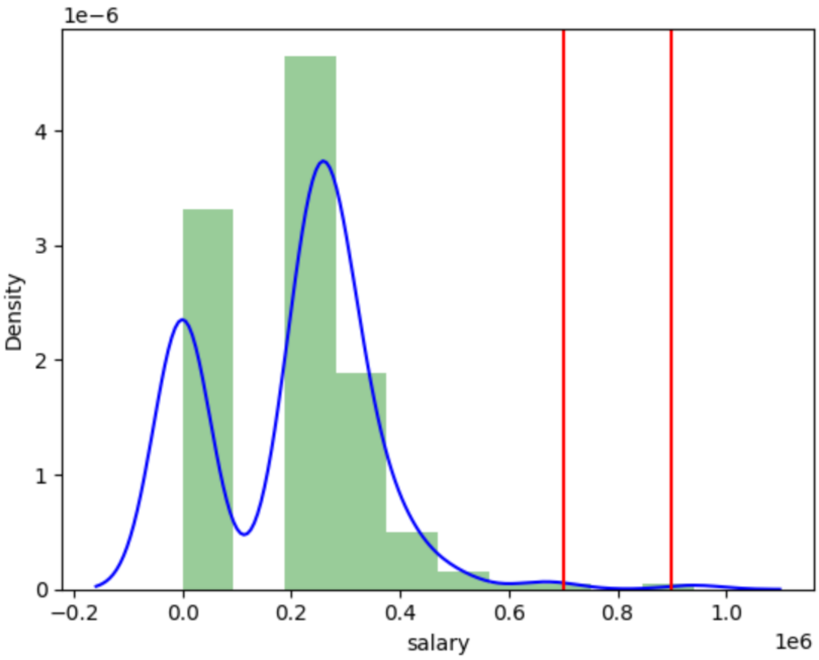
1.11 10) What is the probability Density Function of the salary range from 700000 to 900000?



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

```
Mean=198702.326, Standard Deviation=154780.927
The area between range(700000,900000):0.0005973310593974868

0.0005973310593974868
```



A Probability Density Function (PDF) describes the likelihood of a continuous random variable taking on a particular value. Instead of giving exact probabilities, the PDF shows the density of probability over a range of values.

1.12 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)

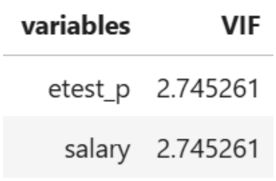
It is found using Paired T-Test – Dependent sample

Output: statistic=5.0049844583693615, pvalue=5.517920600505392e-06, df=58

- Acc to condition p value is 5.51 > 0.05 so accept Null Hypothesis and reject Alternate Hypothesis
- There is no significant difference between etest_p and mba_p

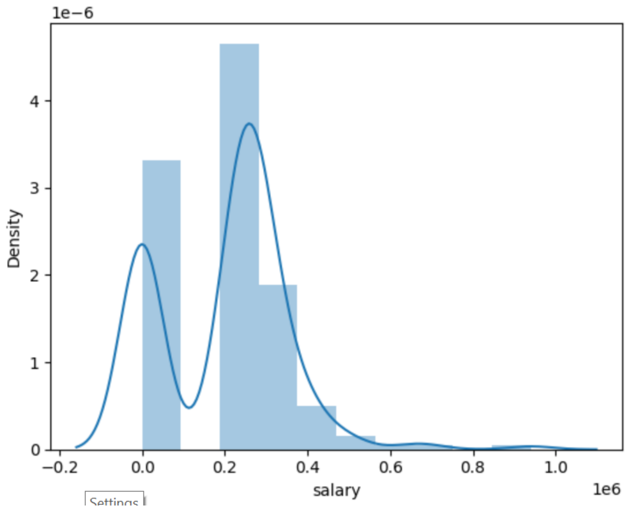
1.13 12) Which parameter is highly correlated with salary?

Correlation if found using Variation Inflation Factor and etest_p is highly correlated with salary.



variables	VIF
etest_p	2.745261
salary	2.745261

1.14 13) Plot any useful graph and explain it.



- This given plot shows normal distribution.
- The salary data is not normally distributed (not a bell curve).
- There are two salary groups (one with low salaries and another with slightly higher salaries).
- The right skewness(positive skewness) suggests a small number of high salaries while most are on the lower end.
- Possible presence of outliers (very high salaries).