Assignment - Inferential Analysis

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

The dataset has NaN values exclusively in the 'salary' column for students with 'status' = "Not Placed". These were replaced with 0.

Justification: Students who are "Not Placed" do not receive a salary, so 0 is a logical and accurate replacement. This avoids distorting analyses (e.g., correlations or means) by treating non-placement as zero income rather than missing data. It also aligns with common practices in placement datasets where salary is only applicable for placed students. After replacement, there are no NaNs left in the dataset.

2. How many of them are not placed?

67 students are not placed (based on 'status' = "Not Placed").

3. Find the reason for non placement from the dataset?

Based on comparative analysis of placed vs. not placed students:

- **Key Insight**: Non-placement is strongly associated with lower academic performance across multiple stages of education. Here's a summary of averages:

Metric	Placed (n=148)	Not Placed (n=67)	Difference
SSC %	71.72	57.54	-14.18
HSC %	69.93	58.40	-11.53
Degree %	68.74	58.85	-9.89
E-test %	72.10	69.59	-2.51
MBA %	62.74	61.61	-1.13

Reasons:

- **Poor early academic performance**: The largest gaps are in SSC, HSC, and degree percentages, suggesting that foundational education scores are a major barrier. Companies likely use these as screening criteria.
- **Slightly lower test and MBA scores**: E-test and MBA percentages are also lower, but the difference is smaller, indicating these are secondary factors.
- **Other factors**: Work experience helps (64% of placed have it vs. 37% of not placed), and specializations like Mkt&Fin have higher placement rates. No strong gender or board bias observed.
- This is inferred from group averages and correlations (e.g., SSC % correlates 0.61 with placement status when encoded as 1/0).

Non-placement isn't due to a single reason but a combination, primarily academic thresholds.

4. What kind of relation between salary and mba_p

There is a weak positive correlation between salary and MBA percentage (mba_p): **r = 0.14**.

- **Interpretation**: Higher MBA scores are slightly associated with higher salaries, but the relationship is not strong (explains only ~2% of variance). This suggests MBA performance has limited influence on salary offers, possibly overshadowed by other factors like prior academics or work experience.
- **Scatterplot** (if visualized) shows a mild upward trend, but with high variance—many high mba_p students have moderate salaries.

5. Which specialization is getting minimum salary?

The specialization getting the minimum average salary is **Mkt&HR** (Marketing & HR).

- **Average Salaries**:
  - Mkt&Fin: ₹288,114
  - Mkt&HR: ₹172,362
- **Justification**: Among placed students, Mkt&HR has consistently lower salaries (min: ₹200,000; max: ₹400,000) compared to Mkt&Fin (min: ₹200,000; max: ₹940,000). This likely reflects market demand—finance roles pay more than HR.

6. How many of them getting above 500000 salary?

6 students are getting above ₹500,000 salary.

- Salaries above 500k: ₹940,000, ₹690,000, ₹650,000, ₹500,000 (two instances), ₹500,000.

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

Hypothesis:

- H_0: Mean of etest_p = Mean of mba_p (no significant difference).
- H_1: Means differ.

Test: One-way ANOVA (F-test) since comparing two continuous variables' distributions.

Results: F-statistic = 98.64, p-value = 4.67e-21 (very small, < 0.05).

Decision: Reject H_0. There is a significant difference between etest_p (mean: 71.47) and mba_p (mean: 62.28). E-test scores are generally higher, possibly due to different assessment natures.

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)

Hypothesis:

- H_0: Mean salary for degree_t = “Sci&Tech” = Mean salary for specialisation = “Mkt&HR” (similar).
- H_1: Means differ.

Test: Independent t-test (unpaired, assuming unequal variances).

Results: t-statistic = 3.20, p-value = 0.0018 (< 0.05).

Decision: Reject H_0. Salaries are not similar. Sci&Tech graduates earn more on average (₹306,977) than Mkt&HR specialists (₹172,362), likely due to technical skills commanding higher pay.

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

Assuming salaries follow a normal distribution, we standardize to z-scores:

- Original salary: Mean (μ) = ₹197,802, Standard Deviation (σ) = ₹154,039.
- Standardized (z-score): z = (salary - μ) / σ.
- Resulting distribution: Mean = 0, Standard Deviation = 1.

Example for first few rows:

- Salary 270,000 → z ≈ 0.47
- Salary 200,000 → z ≈ 0.02
- Salary 0 (not placed) → z ≈ -1.28

This transformation centers the data for easier comparison (e.g., how many SDs above/below mean).

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

Assuming salaries are normally distributed (mean ₹197,802, SD ₹154,039), the probability (integral of PDF over the range) is the cumulative probability:

Probability (P(700,000 ≤ salary ≤ 900,000)) = 0.00084 (or ~0.084%).

- This is very low, as most salaries are below ₹400,000 (only 3 salaries in this range in the data).
- PDF formula (for reference): f(x) = (1 / (σ√(2π))) * e^(-((x-μ)^2)/(2σ^2)).
- Note: Actual data is right-skewed, so normal assumption approximates; empirical probability is ~1.4% (3/215).

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)

Hypothesis (for Sci&Tech students only, n=59):

- H_0: Mean etest_p = Mean mba_p (similar).
- H_1: Means differ.

Test: Paired t-test (since same students' scores).

Results: t-statistic = 4.49, p-value = 1.65e-05 (< 0.05).

Decision: Reject H_0. For Sci&Tech students, etest_p (mean: 73.80) and mba_p (mean: 63.59) are not similar—e-test scores are significantly higher, possibly due to technical aptitude.

12. Which parameter is highly correlated with salary?

The parameter most highly correlated with salary is **ssc_p** (SSC percentage), with r = 0.54.

- **All Correlations with Salary** (among numeric features, excluding self):
    - ssc_p: 0.54 (strongest)
    - hsc_p: 0.45
    - degree_p: 0.40
    - etest_p: 0.18
    - mba_p: 0.14
- Interpretation: Higher secondary school performance (ssc_p) best predicts salary, explaining ~29% of variance. This suggests early academics influence job offers significantly.

13. Plot any useful graph and explain it.

A useful graph is a **pairplot (scatterplot matrix)** of numeric features (ssc_p, hsc_p, degree_p, etest_p, mba_p, salary), colored by placement status.

- **Explanation**: This grid of scatterplots shows relationships between variables. Key observations:
- Strong positive correlations between early academics (ssc_p, hsc_p, degree_p) and salary/placement—higher scores cluster in "Placed" with higher salaries.
- Weaker links for etest_p and mba_p, with more scatter.
- Outliers (e.g., high salaries) often have high scores across boards.
- It highlights that placed students form a "upper cluster" in most pairs, explaining non-placement reasons (Q3).
- **Why useful**: Reveals multicollinearity (e.g., ssc_p and degree_p correlate 0.51) and supports correlation findings (Q12).