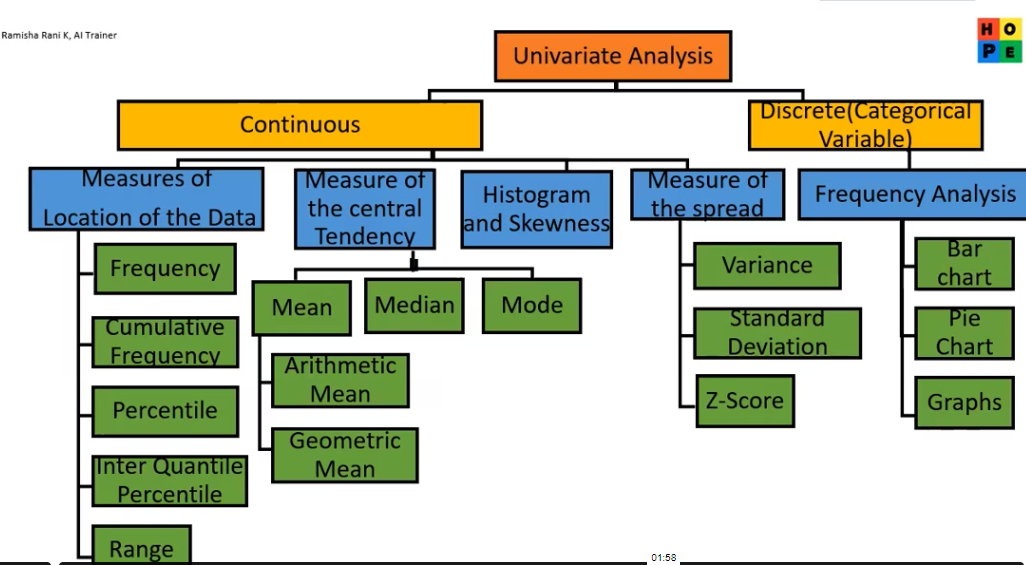
* Dataset.info() -> to detect null and datatypes
* In Univariate analysis:
  + Quant data is used for measure of location of data, central tendency, histogram, skewness and spread of data
  + Qual data is used for frequency analysis – bar/pie/graph



* Measure of central tendency
  + Outlier is not considered by Median. Mean will include the outlier.
  + Mode is repetitions.
* Measure of location of data
  + Percentile – value within a range( intial – Q1 – Q2/Median – Q3/75 – highest value)
* Data preprocessing involves 2 parts before moving to ML:
  + PART 1 – THERE SHOULD BE NO NULL VALUES
  + PART 2 – YOU NEED TO REMOVE THE OUTLIERS

QUESTIONS

1 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.

There are 4 options to replace missing values:

#1)Replace the missing value with central tendency-Mean Median Mode

#2)Entire Row delete

#3)REspective to the problem ,have to replace the missing values

#4)Consider this as a semi supervised, create a model, so that we can predict the missing values.

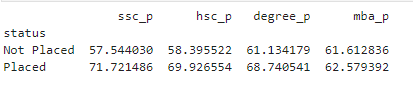
On analyzing the data we find that all the 'not placed' students have an empty salary column, which is logical, so we add zero.

1.3 2)How many of them are not placed?

67 students are not placed.

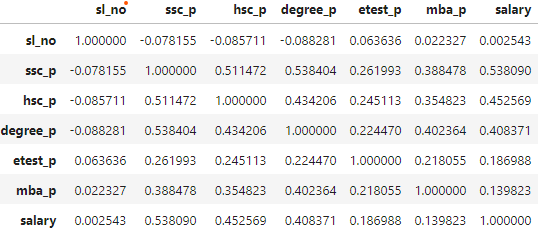
1.4 3)Find the reason for non placement from the dataset?

The non-placement can be due to students scoring lower in their 10th and 12th board exams.



1.5 4 )What kind of relation between salary and mba\_p

Both correlation and covariance are mathematical concepts that require numeric data to compute. We select only numeric values and create a new table. A correlation coefficient of 0.13 between salary and MBA degree percentage (mba\_p) indicates a **weak positive linear relationship** between these two variables.



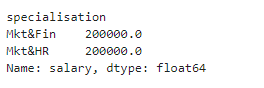
1.6 5)Which specialization is getting minimum salary?

NON-OPTIMIZED CODE

# Step 1: Exclude rows where the salary is zero filtered\_dataset = dataset[dataset['salary'] > 0] # Step 2: Group by 'specialisation' and find the minimum salary for each specialization min\_salary\_by\_specialisation = filtered\_dataset.groupby('specialisation')['salary'].min() # Step 3: Identify the specialization with the minimum salary (excluding zero) min\_specialisation = min\_salary\_by\_specialisation.idxmin() min\_salary = min\_salary\_by\_specialisation.min()

OPTIMIZED CODE

min\_salary\_by\_specialisation = dataset.loc[dataset['salary'] > 0].groupby('specialisation')['salary'].min()



**Explanation:**

* **dataset.loc[dataset['salary'] > 0]**: This uses .loc to filter rows where the salary is greater than zero while maintaining indexing.
* **groupby('specialisation')['salary'].min()**: This groups the filtered data by specialisation and calculates the minimum salary for each group.
* **idxmin() and min()**: These functions identify the specialization with the lowest non-zero minimum salary and get the corresponding salary value.

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

3

1.8 7)Test the Analysis of Variance between etest\_p and mba\_p at signifance level 5%.(Make decision using Hypothesis Testing)

One way ANOVA test gives a P-value =np.float64(2.1814705840542897e-24))

* Since the p-value (2.1814705840542897e-24) is much smaller than 0.05, we can reject the null hypothesis.
* So, there is a statistically significant correlation between mba\_p and the entrance exam scores (etest\_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)

Since we are dealing with two categorical independent variables and one continuous dependent variable (salary), the appropriate test is a **Two-Way ANOVA**.

**Effect of specialization on salary**:

* **P-Value**: 0.000051
* **Interpretation**: The p-value for specialisation is much smaller than 0.05, so we reject the null hypothesis for specialisation. This indicates that there is a statistically significant effect of specialisation on salary. The difference in salaries across different specialisations is significant.

Effect of degree\_t on salary:

P-Value: 0.201122

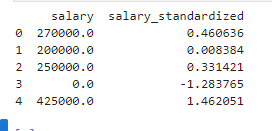
Interpretation: The p-value for degree\_t is greater than 0.05, so we fail to reject the null hypothesis for degree\_t. This suggests that degree\_t alone does not have a statistically significant effect on salary.

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

**Steps to Convert to Standard Normal Distribution:**

1. **Calculate the Mean and Standard Deviation**: Compute the mean and standard deviation of the salary column.
2. **Standardize the Values**: Use the formula for standardization:





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

Mean=198702.326, Standard Deviation=154780.927

The computed probability or likelihood that a salary falls between 7 L and 9 L is very low 0.000597, or 0.0597%.

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)

**Categorical Variable**:

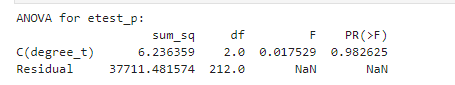
* degree\_t is a categorical variable (e.g., Sci&Tech).

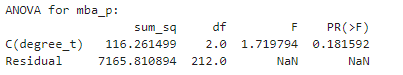
**Continuous Variables**:

* etest\_p (Entrance Test Percentage) and mba\_p (MBA Percentage) are continuous variables.

**Applying the Test:**

Since you want to test the similarity of degree\_t with respect to etest\_p and mba\_p, you are looking at the effect of degree\_t on both etest\_p and mba\_p individually. Therefore, you should use **One-Way ANOVA** separately for each continuous variable.





so the inference is as both p values is above 5%, we accept the null hypothesis that there is no correlation between degree and entrance test marks and no correlation between degree and MBA marks

In other words the degree does not have a statistically significant effect on the entrance test marks and the MBA marks.

1.13 12)Which parameter is highly correlated with salary?

**Steps to Identify the Most Correlated Parameter with salary:**

1. **Extract Correlation Values**:
   * Find the correlation coefficients between salary and other parameters.
2. **Identify Highest Correlation**:
   * Look for the highest absolute value in the correlation column for salary (excluding the correlation of salary with itself, which is always 1).

**Correlation with salary:**

* **ssc\_p**: 0.538090
* **hsc\_p**: 0.452569
* **degree\_p**: 0.408371
* **etest\_p**: 0.186988
* **mba\_p**: 0.139823

**Conclusion:**

Based on the provided correlation matrix:

* **ssc\_p** has the highest positive correlation with salary at 0.538090.
* **hsc\_p** follows with a correlation of 0.452569.
* **degree\_p** has a correlation of 0.408371.

Thus, **ssc\_p** is the parameter that has the highest correlation to salary.

1.14 13) plot any useful graph and explain it. Record Video and Explain each and every concept how you got the answer.

**Summary**

* **Scatter Plot**: Visualizes relationships between two continuous variables.
* **Histogram**: Shows the distribution of a single variable.
* **Box Plot**: Displays spread and identifies outliers in data.
* **Pair Plot**: Examines relationships between multiple variables.
* **Heatmap**: Provides a quick overview of correlations between all numerical variables.

These plots will help you understand the distribution, relationships, and patterns in your dataset, leading to better insights and informed decisions.