BIVARIATE ANALYSIS

COVARIANCE:

To analysis the variance/ differences between two columns or variables:

```
[10]: dataset.cov(numeric_only=True)
```

[10]:

	sl_no	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
sl_no	3870.000000	-52.641355	-59.598879	-41.465047	52.556168	8.102336	1.138318e+04
ssc_p	-52.641355	117.228377	58.853253	42.702550	37.659225	24.535952	9.088585e+05
hsc_p	-59.598879	58.853253	112.063731	33.684453	33.838355	21.517688	7.310079e+05
degree_p	-41.465047	42.702550	33.684453	53.604710	22.078774	17.185200	4.663363e+05
etest_p	52.556168	37.659225	33.838355	22.078774	176.251018	16.886973	3.727004e+05
mba_p	8.102336	24.535952	21.517688	17.185200	16.886973	34.028376	1.239934e+05
salary	11383.177570	908858.485818	731007.850848	466336.264888	372700.449468	123993.387361	2.259185e+10

- Compare *mba_p* and *etest_p* the difference between them is 16.886973 is large positive covariance.
- Compare *etest_p* and *degree_p* the difference between them is 22.0787774 is large positive covariance.
- Compare ssc_p and hsc_*p* the difference between them is 58.853253 is large positive covariance.

Conclusion, mostly in this dataset we have only **large positive covariance**.

CORRELATION:

To find the relation between two columns or variables:

```
]: dataset.corr(numeric_only=True)
```

]:

	sl_no	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
sl_no	1.000000	-0.078155	-0.090500	-0.091039	0.063636	0.022327	0.001217
ssc_p	-0.078155	1.000000	0.513478	0.538686	0.261993	0.388478	0.558475
hsc_p	-0.090500	0.513478	1.000000	0.434606	0.240775	0.348452	0.459424
degree_p	-0.091039	0.538686	0.434606	1.000000	0.227147	0.402376	0.423762
etest_p	0.063636	0.261993	0.240775	0.227147	1.000000	0.218055	0.186775
mba_p	0.022327	0.388478	0.348452	0.402376	0.218055	1.000000	0.141417
salary	0.001217	0.558475	0.459424	0.423762	0.186775	0.141417	1.000000

- Compare *mba_p* and *etest_p* the relation between them is 0.218055 is positive correlation but not much it's just 21% so we consider as zero correlation.
- Compare *etest_p* and *degree_p* the relation between them is 0.227147 is positive correlation but not much it's just 22% so we consider as zero correlation.
- Compare *mba_p* and *salary* the relation between them is 0.14147 is positive correlation but not much it's just 14% so we consider as zero correlation.
- Compare *ssc_p* and *salary* the relation between them is 0.558475 is positive correlation but it very low near to 0.54 means it is low degree of positive correlation

Conclusion, mostly in this dataset we have **zero correlation** and **low degree positive correlation**.

MULTICOLINERARITY:

Target variable should not be linear more than one input.

In the dataset multicollinearity should not present, if it present the prediction will not be accurate.

To overcome multicolineratiy we are using **variance Inflation factor (VIF).**

VIF estimates how much variance is inflated due to multicollinearity in the model.

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
def calc_vif(X):
    vif=pd.DataFrame()
    vif["variables"]=X.columns
    vif["VIF"]=[variance_inflation_factor(X.values, i)for i in range(X.shape[1])]
    return(vif)
```

Above function is used to calculate VIF

vif["variables"]=X.columns

- Here what we are going to give as input or column name that will act as variables here.

vif["VIF"]=[variance_inflation_factor(X.values, i)for i in range(X.shape[1])]

- In this line of code first forloop get executed.
- X.shape is defined as number of columns and rows, X.shape[1] is specified the column which is give as input.
- (X.values, i), it will take the values of the columns.
- variance_inflation_factor(X.values, i), after taking the values from the specific column it will find the VIF and store it in vif.

HOMOSCEDASTICITY AND HETEROSCEDASTICITY:

Homoscedasticity:

- This is assumption of **same variance** is central to linear regression model.
- If we check the error value for the model (actual value – prediction value), that error values are in the linear pattern or in certain interval it will follow the same pattern is called as Homoscedasticity.
- Pattern of amplitude is same in Homoscedasticity.

Heteroscedasticity:

- This is assumption of **different variance** is central to linear regression model.
- When we plot the error value it is scattered the error difference is called as Heteroscedasticity.
- Pattern of amplitude is different in Heteroscedasticity.

Aspect	**Homoscedasticity**	**Heteroscedasticity**
Meaning	Variance of residuals is constant across all levels of independent variables.	Variance of residuals changes (increases or decreases) with the level of independent variables.
Regression Assumption	Satisfies one of the key assumptions of linear regression.	Violates the constant variance assumption of linear regression.
Residual Plot Appearance	Residuals are randomly scattered with no visible pattern.	Residuals form a pattern — often funnel or cone-shaped.
Error Variance	Constant (homogeneous errors).	Non-constant (heterogeneous errors).
Impact on Coefficients	Coefficients remain efficient and unbiased.	Coefficients remain unbiased but become inefficient.
Impact on Standard Errors	Standard errors are reliable and accurate.	Standard errors are biased — leading to wrong p-values and confidence intervals.
Model Reliability	Results are trustworthy; hypothesis tests valid.	Results may be misleading due to unreliable standard errors.
Detection Methods	Visual check: residual plot shows random scatter.	Residual plot shows patterns; Breusch–Pagan or White's test detects it.
Possible Causes	Proper model specification and consistent data spread.	Wide range of data, missing variables, non-linearity, or outliers.
Fix / Remedies	None needed.	Transform variables, use Weighted Least Squares, or robust standard errors.

Comparatively homoscedasticity is best.

T-TEST

It is used to find the similarity between two groups or columns based on mean/probability.

- Paired T-Test (Dependent Sample)
  Same group and different condition.
- Unpaired T-Test (Independent Sample)
  Different group and same condition.

Up to 5% pvalue is acceptable.

HYPOTHESIS TESTING

It means if we accept the statement or reject it is the concept.

- Null hypothesis.
- Alternate hypothesis.

Null hypothesis:

The statistic is not same for both sample and population.

Alternate hypothesis:

The statistic is same for both sample and population.

If pvalue <0.05% reject the null hypothesis and accept alternate hypothesis.

ANOVA [ANALYSIS OF VARIANCE]

Anova is a statistical method used to compare the means of three or more groups to see there is a significant difference between them.

Why Anova?

Sometimes, multiple T-test increase the chance of error, but anova solves this by testing all groups together in one test.

Types:

- One-way classification.
  It will used one independent variable(factor) which his compared with dependent variable.

- Two-way classification.
  It will used two independent variables(factors) which is compared With dependent variable.