1. Co - Variance & Correlation:

In [5]:
```
dataset.cov()
```
Out[5]:

	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

In [6]:
```
dataset.corr()
```
Out[6]:

	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

Inference Table:

Description	**Covariance Between**		**Correlation Between**
	Degree pass mark & E-test pass mark	E-test pass mark & MBA pass mark	MBA pass mark & Salary
Type	Positive Covariance	Positive Covariance	Positive Correlation
Difference or Related By Quantity	22.08	16.89	0.14
Level	**Differing Level**		**Correlation Level**
	Small	Small	Very Small
Take away	Degree pass mark & E-test pass mark varies by a small amount	E-test pass mark & MBA pass mark varies by a small amount	MBA pass mark & Salary exhibits a very small correlation

2.VIF Code Explanation:

VIF Code Explanation:

In [11]:

```python
#Importing necessary Library
from statsmodels.stats.outliers_influence import variance_inflation_factor


#Creating a function to calcilate VIF
def calc_vif(X):

    # To Calculate VIF


    #Creating a table under the variable vif
    vif = pd.DataFrame()

    # Assigning input coloumn under X to act as variable
    vif["variables"] = X.columns

    # Creating a for loop for the input coloumns & calculating VIF for the values under that column
    vif["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]

    # Returns back the table
    return(vif)

```

3.Ways to Remove Multicollinearity:

Methods to remove multicollinearity:

► Re specification of the model

► Use of additional data/collect more new data

► Independent estimation of parameters

► Principal Component Regression

► Ridge Regression & Lasso Regression

Model re specification :

► Multicollinearity in most cases is caused by the high correlation between two predictors. In such situation a possible remedy could be restructuring the predictors.

► For example, if x_1, x_2 and x_3 are nearly linearly related to each other, then it might be possible to find a function of these predictors such as $x = (x_1+x_2)/x_3$ or $x = x_1x_2x_3$ such that the information is preserved and also multicollinearity is reduced.

► Another popularly used method of model re specification is variable elimination.

►Removing highly correlated variables.

► This method is usually very effective. However, removal of a predictor may also indicate that the predictive power of the model has been compromised.

Use of additional data/collect more data :

► Multicollinearity is a sampling phenomenon.
Thus it is possible that in another sample with data for the same variables collinearity may not be so serious.

► Including more data i.e. increasing the sample size is also a good option. Since the standard deviation of the parameter is a function of sample size, increasing the sample size will decrease the standard deviation.

► But collecting additional data is not always possible because of economic constraints.

Independent estimation of parameters:

► Let us consider the Ando-Modigliani consumption equation given by;

$$C_t = \beta_0 + \beta_1 + \beta_2 W_t + u_t$$

For the given case we can obtain the estimate of β_1 say β_1'. Then the above model can be written as

$$C_t - \beta_1' W_t = \beta_0 + \beta_2 W_t + u_t$$

and now this becomes our new problem.

► Thus treating β_1'as known values will help in estimating β_2 more precisely.

► To handle such problem we use mixed estimation approaches.

Principle component regression:

► Principal component regression is a technique employed to fit multiple regression model, where the assumption of multicollinearity is violated.

► In case of multicollinearity, although the parameter estimates are unbiased their standard deviations become very high. In principle component regression we add a degree of bias to the regression coefficient estimates, and achieve a lower standard deviation.

4. Purpose of Homo & Heteroscedasticity:

Significance of Homoscedasticity:

Homoscedasticity is a key assumption for linear regression. Data that exhibits homoscedasticity is appropriate for linear regression. Violating homoscedasticity means that the dataset will need to be transformed or changed in some way, or a different model selected (e.g. WOLS instead of OLS). It's worth noting that OLS can tolerate some heteroscedasticity.

Though, it’s important also to note that OLS regression can tolerate some heteroscedasticity,one rule of thumb suggests that “the highest variability shouldn’t be greater than four times that of the smallest.”

This is an important assumption of parametric statistical tests because they are sensitive to any dissimilarities. Uneven variances in samples result in biased and skewed test results.

Significance of Heteroscedasticity:

The concept of heteroscedasticity - the opposite being homoscedasticity - is used in statistics, especially in the context of linear regression or for time series analysis, to describe the case where the variance of errors of the model is not the same for all observations, while often one of the basic assumption in modelling is that the variances are homogeneous and that the errors of the model are identically distributed.

In linear regression analysis, the fact that the errors of the model (also named residuals) are not homoscedastic has the consequence that the model coefficients estimated using ordinary least squares (OLS) are neither unbiased nor those with minimum variance. The estimation of their variance is not reliable.

5. Paired & Unpaired T-Test:

Independant Sample- Unpaired T Test

Different Groups(Hsc other board & Hsc central board) but same condition(salary)

Different Groups(Science & Commerce) but same condition(MBA_p)

In [13]:
```python
from scipy.stats import ttest_ind
dataset=dataset.dropna()
Others = dataset[dataset['hsc_b']=='Others']['salary']
Central = dataset[dataset['hsc_b']=='Central']['salary']
ttest_ind(Others, Central)
```

Out[13]:
```
Ttest_indResult(statistic=0.30570032095155825, pvalue=0.7601313863865756)
```

In [14]:
```python
from scipy.stats import ttest_ind
dataset=dataset.dropna()
Science = dataset[dataset['hsc_s']=='Science']['mba_p']
Commerce = dataset[dataset['hsc_s']=='Commerce']['mba_p']
ttest_ind(Science, Commerce)
```

Out[14]:
```
Ttest_indResult(statistic=0.7331285580404581, pvalue=0.46432995253854314)
```

Dependant Sample-Paired T_Test

Same Group(Commerce) but Different Conditions(etest_p & mba_p)

In [15]:
```python
from scipy.stats import ttest_rel
#dataset=dataset.dropna()
etest_p = dataset[dataset['hsc_s']=='Commerce']['etest_p']
mba_p= dataset[dataset['hsc_s']=='Commerce']['mba_p']
ttest_rel(etest_p, mba_p)

```

Out[15]:
```
TtestResult(statistic=7.868552092606871, pvalue=2.462926468454984e-12,
df=112)
```

Inference :

Unpaired T-test:

1. Different Groups(Hsc other board & Hsc central board) but same condition(salary)
p- Value= 0.76 > 0.05 -----There exists larger difference in the salary received by the other board & central board students under hsc.

2. Different Groups(Science & Commerce) but same condition(MBA_p)
p- Value = 0.46 > 0.05 ----- There exists larger difference in the mba pass mark obtained by the science & commerce group students.

Paired T- test:

1. Same Group(Commerce) but Different Conditions(etest_p & mba_p)
p- Value= $2.46x\ 10^{-12}$ < 0.05 ----- There exists similarity between the etest_pass mark & mba_pass mark obtained by the commerce group students.

6. ANAVO- 5 Eg. Problem Statements For One- Way Classification & Two Way Classification

One- Way Classification Example Problems:

1. Three processes A, B and C are tested to see whether their outputs are equivalent. The following observations of outputs are made:

A	10	12	13
B	9	11	10
C	11	10	15

Analysis the data to test the significance of the differences between the output in three process.

2. A test was given to three students taken at random from the third class of three schools of a town. The individual scores are

School I	9	7	7
School II	6	5	6
School III	7	4	5

Analysis the data to test the significance of the differences between the price marks in three schools.

3. The following table gives the retail prices of a commodity in (Rs. Per Kg) in some shops selected at random in 3 cities.

City A	22	22	17
City B	16	25	26
City C	27	14	25

Analysis the data to test the significance of the differences between the price of commodity in three cities.

Two- Way Classification Example Problems:

4. The following are the defective pieces produced by four operators working in turn, on four different machines:

Operator

Machine	I	II	III	IV
A	3	2	3	2
B	3	2	3	4
C	2	3	4	3
D	3	4	3	2

Perform analysis of variance at 5% level of significance to ascertain whether variability in production is due to variability in operator's performance or variability in machine's performance.

5. For experiments determine the moisture content of sample of a powder, each man taking a sample from each of three consignments Their assessments are:

Consignment

Observer	1	2	3
1	9	10	9
2	12	11	9
3	11	10	10

Perform an analysis of variance of these data and discuss if there is any significant difference between consignments or between observers.

7. Code For Two Way Classification:

Code:

ANAVO : Analysis of Variance

Two Way Classification

In [*]:

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

# Fit the Two-Way ANOVA model
model = ols('salary ~ C(degree_p) + C(mba_p) + C(degree_p):C(mba_p)', data=dataset).fit()

# Perform the ANOVA
anova_table = sm.stats.anova_lm(model, typ=2)
anova_table
```

--------The End--------

முத்து வசுமதி