In our dataset has 62 rows and 14 columns are presented in survey dataset. The 2 float, 64 int and 6 object is reflected in our dataset.

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Soc Network
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	
1	2	Male	23	Senior	Management	Yes	3.6	Part-Time	25.0	
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	45.0	
3	4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40.0	
4	5	Male	23	Senior	Other	Undecided	2.8	Unemployed	40.0	
...	...	...	...	...	...	...	...	...	...	
57	58	Female	21	Senior	International Business	No	2.4	Part-Time	40.0	
58	59	Female	20	Junior	CIS	No	2.9	Part-Time	40.0	

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62 entries, 0 to 61
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   ID                 62 non-null     int64
 1   Gender             62 non-null     object
 2   Age                62 non-null     int64
 3   Class              62 non-null     object
 4   Major              62 non-null     object
 5   Grad Intention     62 non-null     object
 6   GPA                62 non-null     float64
 7   Employment         62 non-null     object
 8   Salary             62 non-null     float64
```

1.1. What is the probability that a randomly selected CMSU student will be male?

In this Question, first we need to take value counts of male and female.

```
Female    33
Male      29
Name: Gender, dtype: int64
```

Then will get the probability thar random CMSU male student.

```
Probability of selected male student is 46.77
```

1.2. What is the probability that a randomly selected CMSU student will be female?

Similarly we need to value counts and count of female is 33.

```
Probability of selected Female student is 53.23
```

1.3. What is the conditional probability of different majors among the male students in CMSU?

The initial value count of majors has listed below

Retailing/Marketing	14
Economics/Finance	11
Management	10
Other	7
Accounting	7
International Business	6
CIS	4

Out of the total majors of male count of major has 25.

Management	6
Retailing/Marketing	5
Other	4
Economics/Finance	4
Accounting	4
Undecided	3
International Business	2

Probability of management majors among the male students in CMSU is **20.7**

Probability marketing majors among the male students in CMSU is **17.2**

Probability majors among the male students in CMSU is **13.8**

Probability economics majors among the male students in CMSU is **13.8**

Probability accounting majors among the male students in CMSU is **13.8**

Probability undecided majors among the male students in CMSU is **10.299999999999999**

Probability international Business majors among the male students in CMSU is **6.9**

Probability CIS majors among the male students in CMSU is **3.4000000000000004**

1.4. What is the conditional probability of different majors among the female students of CMSU?

Major counts for female.

Retailing/Marketing	9
Economics/Finance	7
Management	4
International Business	4
Other	3
CIS	2

Probability marketing majors among the female students in CMSU is **27.3**

Probability economics majors among the female students in CMSU is **21.21**

Probability management majors among the female students in CMSU is **12.120000000000001**

Probability international Business majors among the female students in CMSU is **12.120000000000001**

Probability other majors among the female students in CMSU is **9.09**

Probability CIS majors among the female students in CMSU is **9.09**

Probability accounting majors among the female students in CMSU is **9.09**

1.5. What is the probability that a randomly chosen student is a male and intends to graduate?

Total categories in

Yes	17
Undecided	9
No	3

Probability that a randomly chosen student of male is yes 58.621

Probability that a randomly chosen student of male is no 10.345

Probability that a randomly chosen student of male is undecided 31.034

1.6. What is the probability that a randomly selected student is a female and does NOT have a laptop?

Not having laptop in female.

Laptop	29
Tablet	2
Desktop	2

Probability of female that not having laptop 12.1

1.7. What is the probability that a randomly chosen student is a male or has full-time employment?

Part-Time	19
Full-Time	7
Unemployed	3

Probability of male in full time employment **24.14.**

1.8. What is the conditional probability that given a female student is randomly chosen, she is majoring in international business or management?

The category of major is

```
Retailing/Marketing        9
Economics/Finance          7
Management                 4
International Business     4
Other                      3
```

The above counts are same in management and international Business.
Probability of female in majoring international business or management 12.1.

1.9. Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think the graduate intention and being female are independent events?

Here is the below table for intent to graduate is yes and no

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Socia Networkin
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1
1	2	Male	23	Senior	Management	Yes	3.6	Part-Time	25.0	1
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	45.0	2
3	4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40.0	4
8	9	Female	20	Junior	Management	Yes	3.6	Unemployed	30.0	0
10	11	Female	23	Senior	Economics/Finance	Yes	2.8	Full-Time	50.0	2

This is a table for female students for grad intention

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Soc Network
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	
8	9	Female	20	Junior	Management	Yes	3.6	Unemployed	30.0	
10	11	Female	23	Senior	Economics/Finance	Yes	2.8	Full-Time	50.0	
24	25	Female	20	Junior	Economics/Finance	Yes	3.0	Part-Time	55.0	
27	28	Female	20	Junior	International Business	Yes	2.9	Part-Time	50.0	
35	36	Female	26	Junior	Accounting	Yes	3.3	Part-Time	60.0	

H0 = The Graduate intention and female are independent events

H1 = The Graduate intention and female are not independent events

So the **P_value (Probabilistic value)is 1.0** so we accept the null hypothesis.
We can conclude grad intention and female are **independent events.**

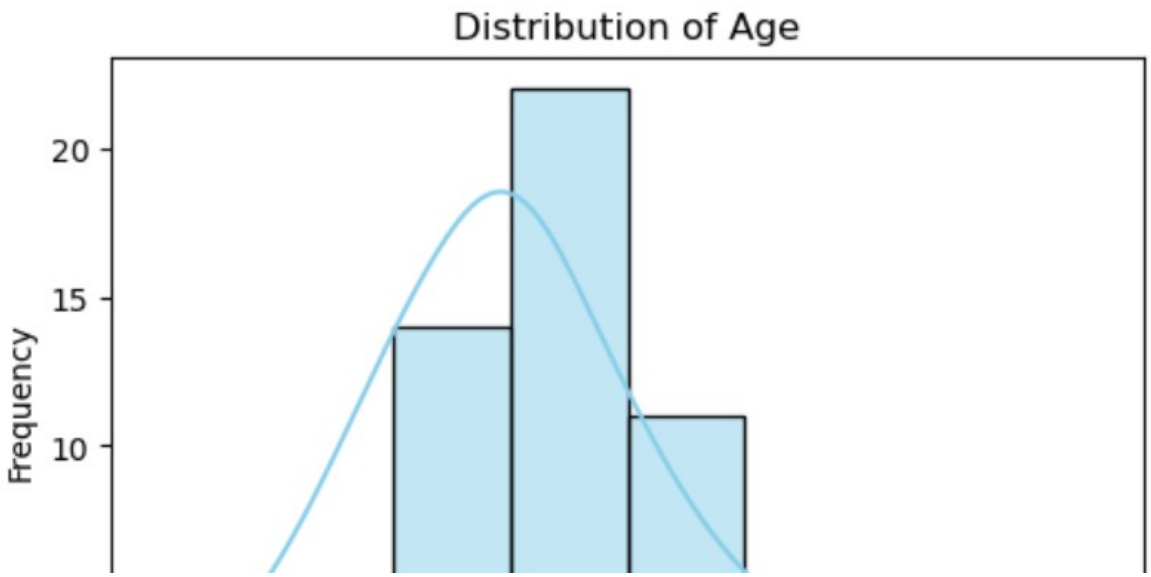
1.10. If a student is chosen randomly, what is the probability that his/her GPA is less than 3?

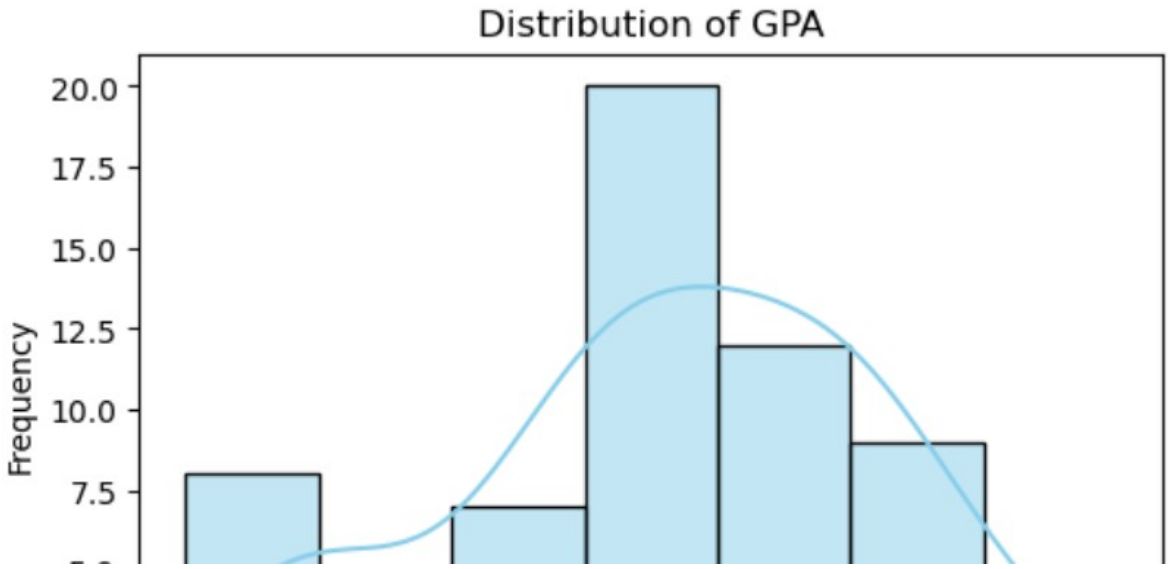
Mean of GPA is **3.12**
The standard deviation is **0.37738**
The probability of less than 3 of his or her GPA is **0.362.**

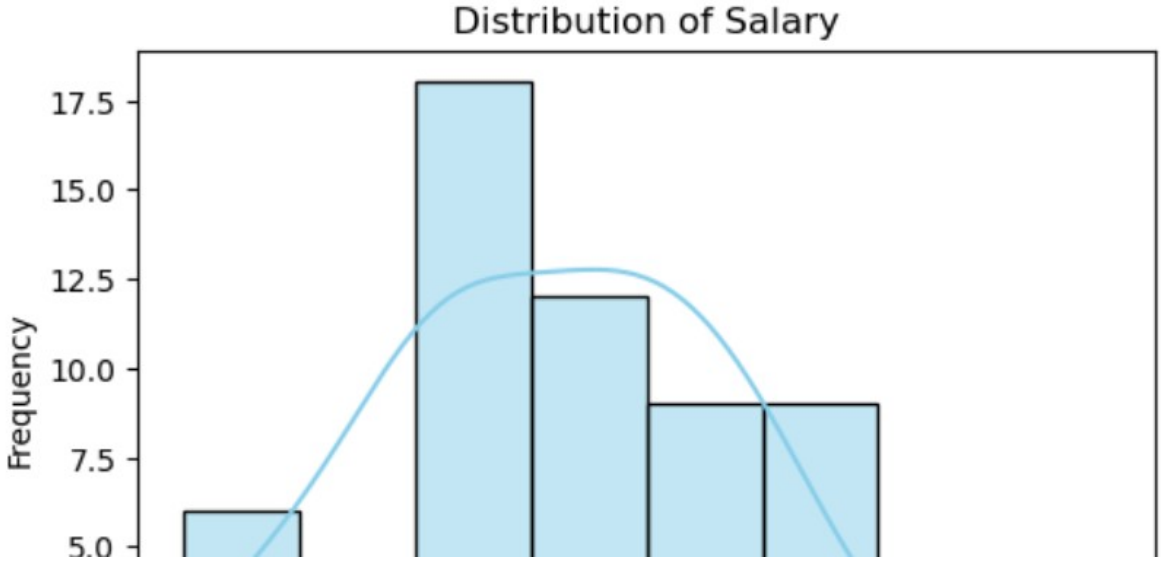
1.11. What is the conditional probability that a randomly selected male earns 50 or more? Find the conditional probability that a randomly selected female earns 50 or more.

Probability of male in earn 50 or more 73.7

Probability of female in earn 50 or more 54.55


Distribution of Age
Frequency
20
15
10


Distribution of GPA
Frequency
20.0
17.5
15.0
12.5
10.0
7.5


Distribution of Salary
Frequency
17.5
15.0
12.5
10.0
7.5
5.0

1.12. For each of the continuous variables in the dataset, comment whether they follow a normal distribution. Write a note summarizing your conclusions.

Inferences:

Yes, as per the histogram it follows a normal distribution

There is a normal distribution occurred in age, Salary and GPA. It seems to be female students is more when compared to male student. The Management major student has a large number among the diverse majors in male category and female is Retailing/Marketing. The high chance of getting into graduation that is high percentage male is interested. The males is more earning when compared to females about of more than 50%.

2.1 Do you think there is evidence that means moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

The P_value is low so we have enough evidence to reject the null hypothesis and we can conclude the both type of shingles are within permissible limits.

Level of significance(Alpha) = 0.05

µA:Population mean moisture content for type A shingles.

µB:Population mean moisture content for type B shingles.

Null Hypothesis = 0.35 = µA = µB Alternate Hypothesis < 0.35 that is µA <0.35 µB<0.35

2.2 Do you think that the population mean for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?

1)Shapiro Wilk Test

In Shapiro wilk test for normality, there is no evidence to reject the null hypothesis. The data appears in normal distribution.

The data meets the assumption of normality for conducting parametric test.

2)Levene's test(Variance test)

The Levene's test result indictes there might be issue with the variance of data that needs to be addressed.

Hypothesis Test:

Alpha = 0.05
Null Hypothesis is population means for **Shingles A and Shingles B** are Equal.
Alternate Hypothesis is population mean for **shingles A and B** are not equal.

Two sample test p_values 0.0

We can conclude means for shingles **A and B are Equal**

Problem 3A:

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [SalaryData.csvView in a new window] are collected and each person's educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor's, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination.

1)State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results. Perform a one-way ANOVA on Salary with respect to Occupation. State whether the null hypothesis is accepted or rejected based on the ANOVA results. If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result. (Non-Graded).

H0 - Null hypothesis - There is no significant difference between the salary across the different level of occupation

Ha - alternative Hypothesis -There is a significant difference between the salary across the different level of occupation

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	0.884144	0.458508

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	30.95628	1.257709e-08
Residual	37.0	6.137256e+10	1.658718e+09	NaN	NaN

P_value is 0.45 and 1.25 for occupation and education respectively. So as per the 95 % of confidence interval (significance level alpha = 0.05), we cannot the null hypothesis.

Occupation and Education are the above significance level (0.05) so the p_value of occupation:education is 2.9137

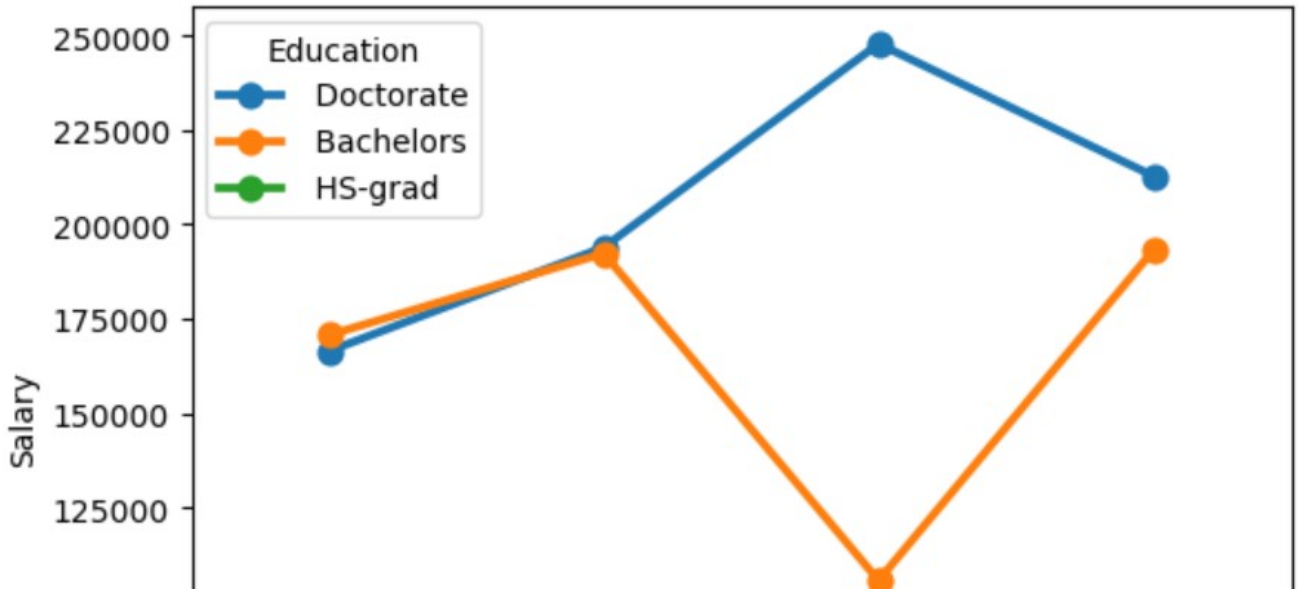
2)Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education*Occupation). State the null and alternative hypotheses and state your results.

Null hypothesis: Salary is hypothesized to depend on educational qualification and occupation.

Alternative hypothesis: Salary is not hypothesized to depend on educational qualification and occupation.

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	5.277862	4.993238e-03
C(Education)	2.0	9.695663e+10	4.847831e+10	68.176603	1.090908e-11
C(Occupation):C(Education)	6.0	3.523330e+10	5.872217e+09	8.258287	2.913740e-05

3)Explain the business implications of performing ANOVA for this particular case study. Please reflect on all that you have learned while working on this project. This step is critical in cementing all your concepts and closing the loop. Please write down your thoughts here.



We can see some interaction in bachelors and doctors until 2 lakh.
Unfortunately we didn't use the **Tukey HSD test** which is used to identify which **class means are differently from each other.**

Occupation and Education are the above significance level (0.05) so the p_value **of occupation:education is 2.9137**

Null hypothesis: Salary is hypothesized to depend on educational qualification and occupation.

Alternative hypothesis: Salary is not hypothesized to depend on educational qualification and occupation.

So we cannot reject the **Null Hypothesis and hence salary is hypothesized to depend on educational qualification and occupation.**