Advance Statistics Project 1,2&3:

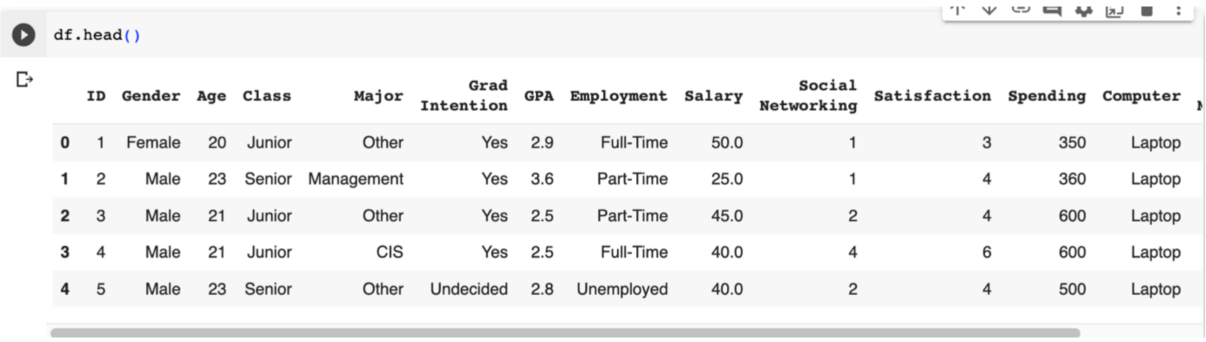
By Ganesh Aryan

The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the ***Survey*** data set).

Loading the Data:

```
df = pd.read_csv('Survey.csv')
```

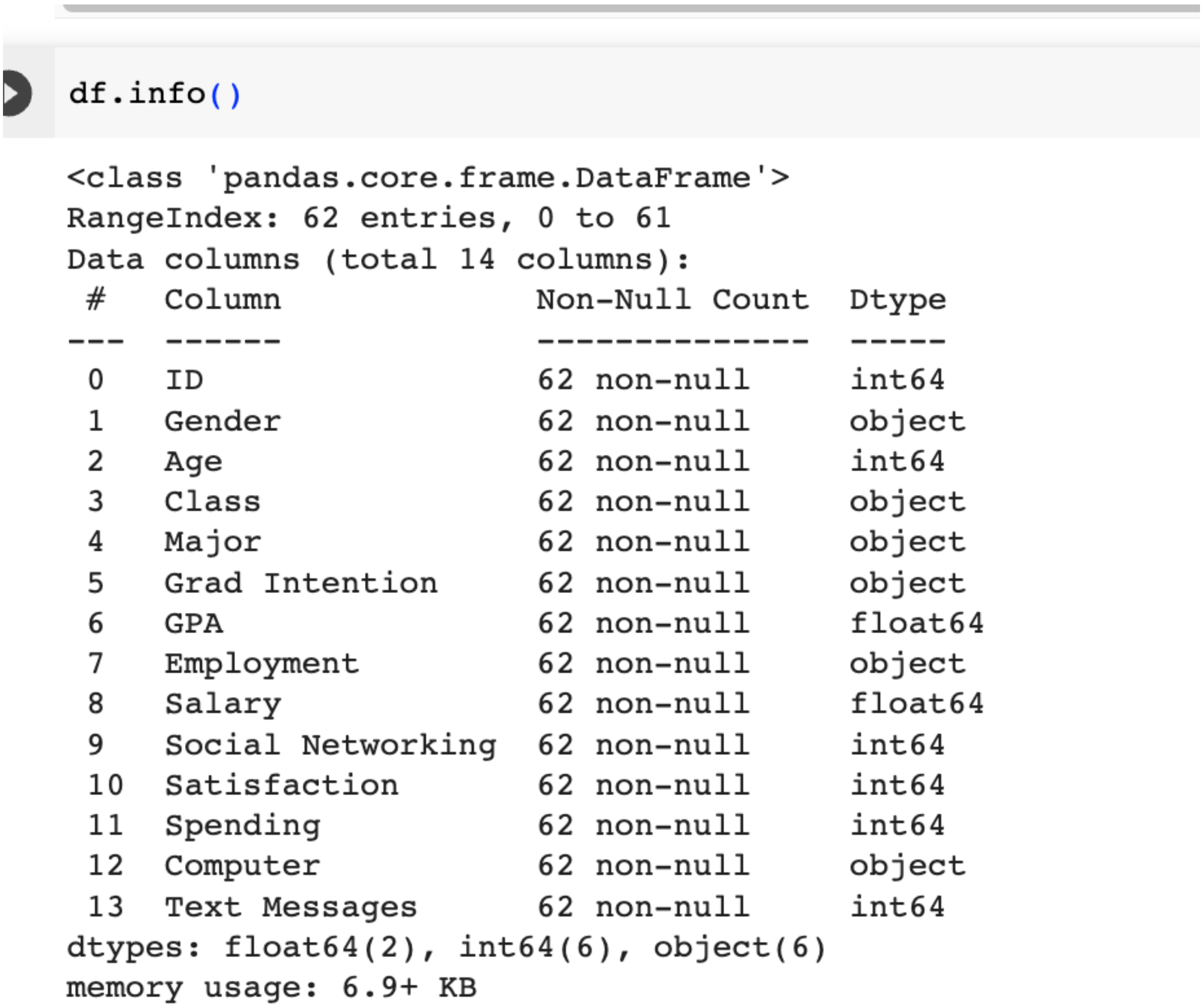
Top 5 rows of the data:



```
df.head()
```

	ID	Gender	Age	Class	Major	Grad Intention	GPA	Employment	Salary	Social Networking	Satisfaction	Spending	Computer
0	1	Female	20	Junior	Other	Yes	2.9	Full-Time	50.0	1	3	350	Laptop
1	2	Male	23	Senior	Management	Yes	3.6	Part-Time	25.0	1	4	360	Laptop
2	3	Male	21	Junior	Other	Yes	2.5	Part-Time	45.0	2	4	600	Laptop
3	4	Male	21	Junior	CIS	Yes	2.5	Full-Time	40.0	4	6	600	Laptop
4	5	Male	23	Senior	Other	Undecided	2.8	Unemployed	40.0	2	4	500	Laptop

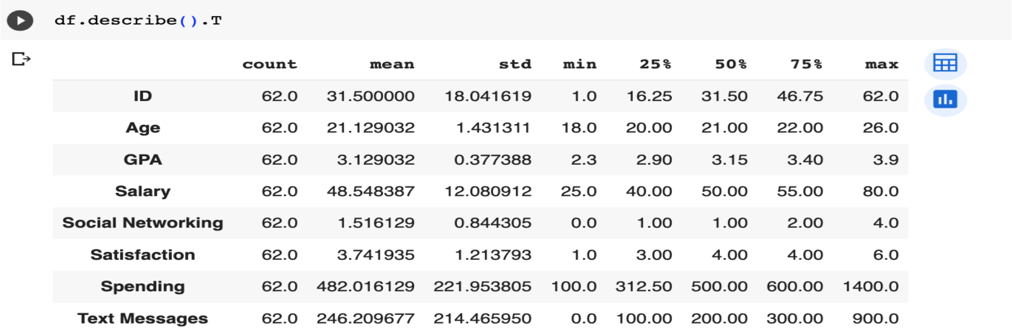
Summary of the data:



```
df.info()
```

```
<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
 9   Social Networking  62 non-null     int64  
 10  Satisfaction       62 non-null     int64  
 11  Spending           62 non-null     int64  
 12  Computer           62 non-null     object 
 13  Text Messages      62 non-null     int64  
dtypes: float64(2), int64(6), object(6)
memory usage: 6.9+ KB
```

Descriptive Analysis:



```
df.describe().T
```

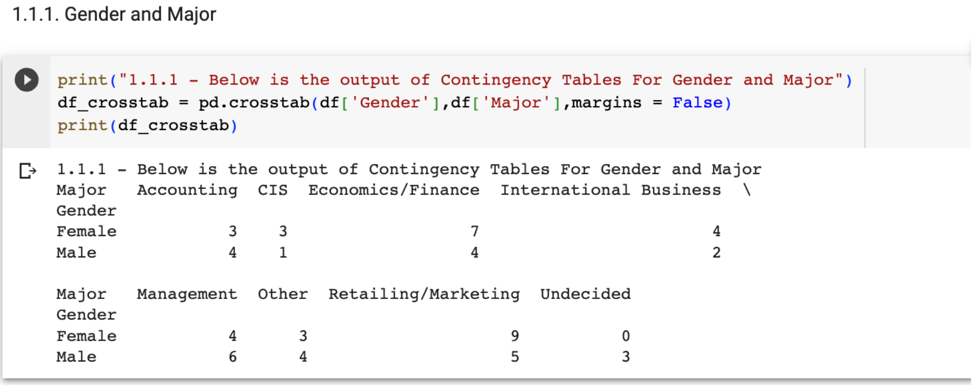
	count	mean	std	min	25%	50%	75%	max
ID	62.0	31.500000	18.041619	1.0	16.25	31.50	46.75	62.0
Age	62.0	21.129032	1.431311	18.0	20.00	21.00	22.00	26.0
GPA	62.0	3.129032	0.377388	2.3	2.90	3.15	3.40	3.9
Salary	62.0	48.548387	12.080912	25.0	40.00	50.00	55.00	80.0
Social Networking	62.0	1.516129	0.844305	0.0	1.00	1.00	2.00	4.0
Satisfaction	62.0	3.741935	1.213793	1.0	3.00	4.00	4.00	6.0
Spending	62.0	482.016129	221.953805	100.0	312.50	500.00	600.00	1400.0
Text Messages	62.0	246.209677	214.465950	0.0	100.00	200.00	300.00	900.0

Checking the missing value

No Null value present in data.

1.1. For this data, construct the following contingency tables (Keep Gender as row variable)

1.1.1. Gender and Major



1.1.1. Gender and Major

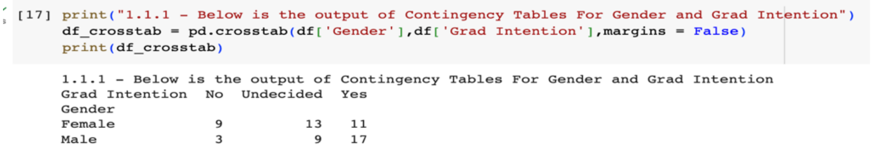
```
print("1.1.1 - Below is the output of Contingency Tables For Gender and Major")
df_crosstab = pd.crosstab(df['Gender'],df['Major'],margins = False)
print(df_crosstab)
```

```
1.1.1 - Below is the output of Contingency Tables For Gender and Major
Major   Accounting  CIS  Economics/Finance  International Business  \
Gender
Female           3    3                  7                       4
Male             4    1                  4                       2

Major   Management  Other  Retailing/Marketing  Undecided
Gender
Female           4      3                    9          0
Male             6      4                    5          3
```

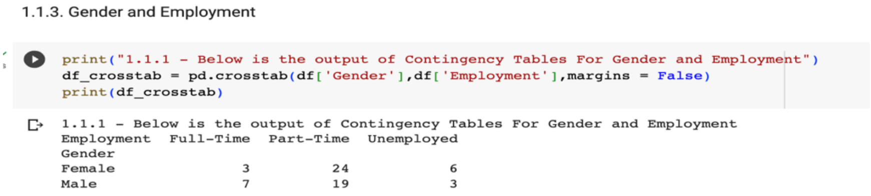
1.1.2. Gender and Grad Intention

1.1.3. Gender and Employment



```
[17] print("1.1.1 - Below is the output of Contingency Tables For Gender and Grad Intention")
     df_crosstab = pd.crosstab(df['Gender'],df['Grad Intention'],margins = False)
     print(df_crosstab)
```

```
1.1.1 - Below is the output of Contingency Tables For Gender and Grad Intention
Grad Intention  No  Undecided  Yes
Gender
Female           9         13   11
Male             3          9   17
```

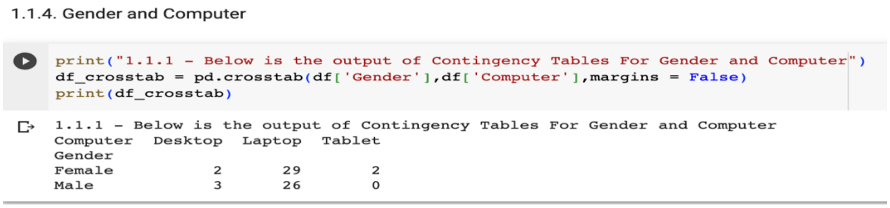


1.1.3. Gender and Employment

```
print("1.1.1 - Below is the output of Contingency Tables For Gender and Employment")
df_crosstab = pd.crosstab(df['Gender'],df['Employment'],margins = False)
print(df_crosstab)
```

```
1.1.1 - Below is the output of Contingency Tables For Gender and Employment
Employment  Full-Time  Part-Time  Unemployed
Gender
Female              3         24           6
Male                7         19           3
```

1.1.4. Gender and Computer



1.1.4. Gender and Computer

```
print("1.1.1 - Below is the output of Contingency Tables For Gender and Computer")
df_crosstab = pd.crosstab(df['Gender'],df['Computer'],margins = False)
print(df_crosstab)
```

```
1.1.1 - Below is the output of Contingency Tables For Gender and Computer
Computer  Desktop  Laptop  Tablet
Gender
Female          2      29       2
Male            3      26       0
```

1.2. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

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

1.2.1 - Probability that a randomly selected CMSU student will be male 0.46774193548387094 (47%)

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

1.2.2 - Probability that a randomly selected CMSU student will be Female 0.532258064516129 (53%)

1.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

1.3.1. Find the conditional probability of different majors among the male students in CMSU.

Probability of Males opting for accounting 13.79%

Probability of Males opting for Economics/Finance 13.79%

Probability of Males opting for CIS 3.44%

Probability of Males opting for International Business 6.89%

Probability of Males opting for Management 20.68%

Probability of Males opting for Other 13.79%

Probability of Males opting for Retailing/Marketing 17.24%

Probability of Males opting for Undecided 10.34%

1.3.2 Find the conditional probability of different majors among the female students of CMSU.

Probability of females opting for accounting 9.09%

Probability of Females opting for Economics/Finance 21.21%

Probability of Males opting for CIS 9.09%

Probability of Females opting for International Business 12.12%

Probability of Females opting for Management 12.12%

Probability of Females opting for Other 9.09%

Probability of Females opting for Retailing/Marketing 27.27%

1.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

1.4.1. Find the probability That a randomly chosen student is a male and intends to graduate.

The probability That a randomly chosen student is a male and intends to graduate 58.62%

1.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop.

The probability that a randomly selected student is a female and does not have a laptop 12.12%

1.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

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

The probability that a randomly chosen student is a male or has a full-time employment 27.70%

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

The conditional probability that given a female student is randomly chosen, she is majoring in international business or Management 16.12%

1.6. 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?

```
#The Probability that a randomly selected student 'being female'
#The probability that a randomly selected student the graduate
intention and being fem
#P(Grad Intention Yes) = 28/40*100 = 70
#P(Grad Intention Yes Female) = 11/20*100 =55
#The Probability are not equal. This suggests that the two events are
independent
```

1.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.

Answer the following questions based on the data

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

27.41%

```
Observation - Using contingency tables of Gender and GPA we got the
total numbers of students and number of students GPA less than 3 and
post calculation we found out that - Probability that student is chosen
randomly and that his/her GPA is less than 3 is 27.41%
```

1.7.2. Find 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.

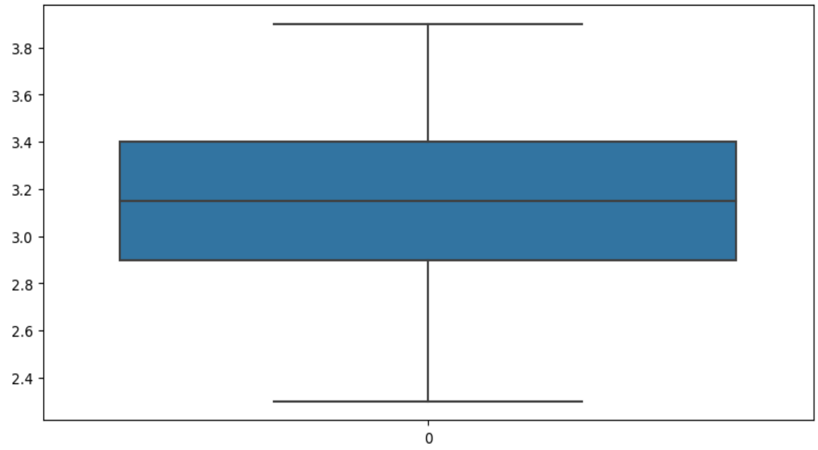
48.27%

```
Probability that randomly selected male earns 50 or more:
48.275862068965516
```

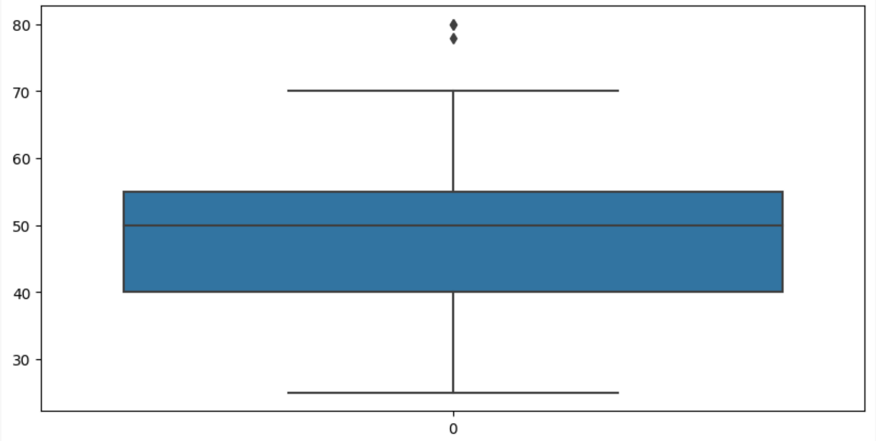
1.8. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions**.**

Box Plot for all Variables:

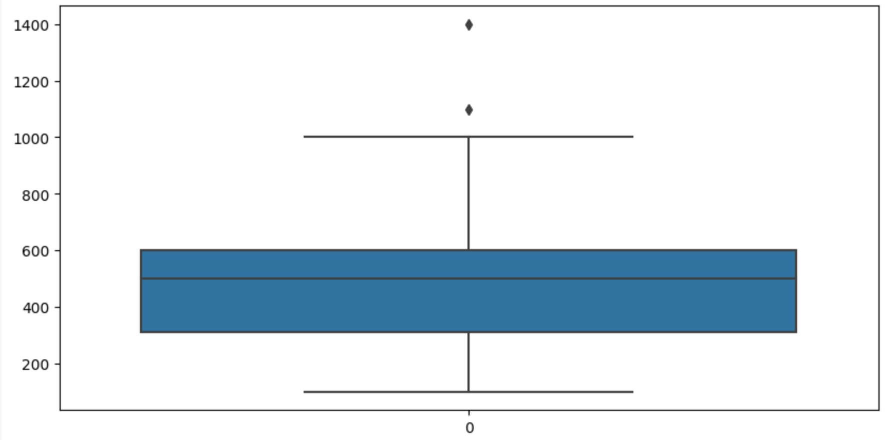
“GPA”



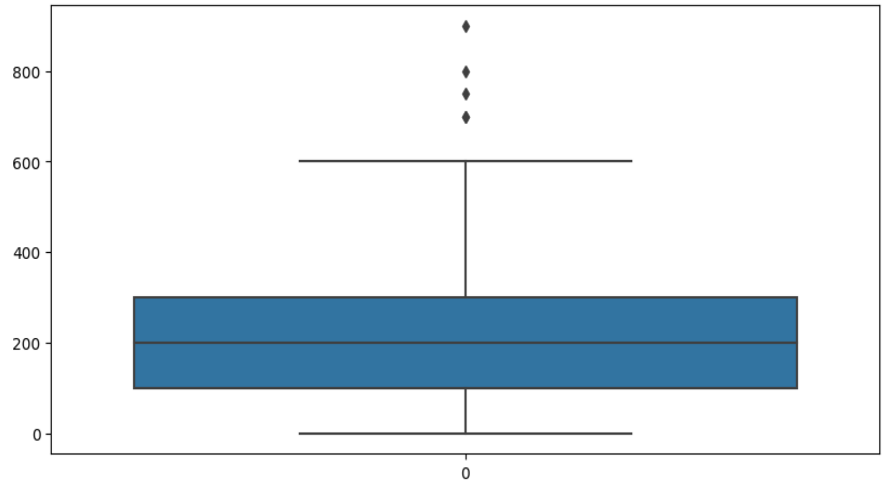
“Salary”



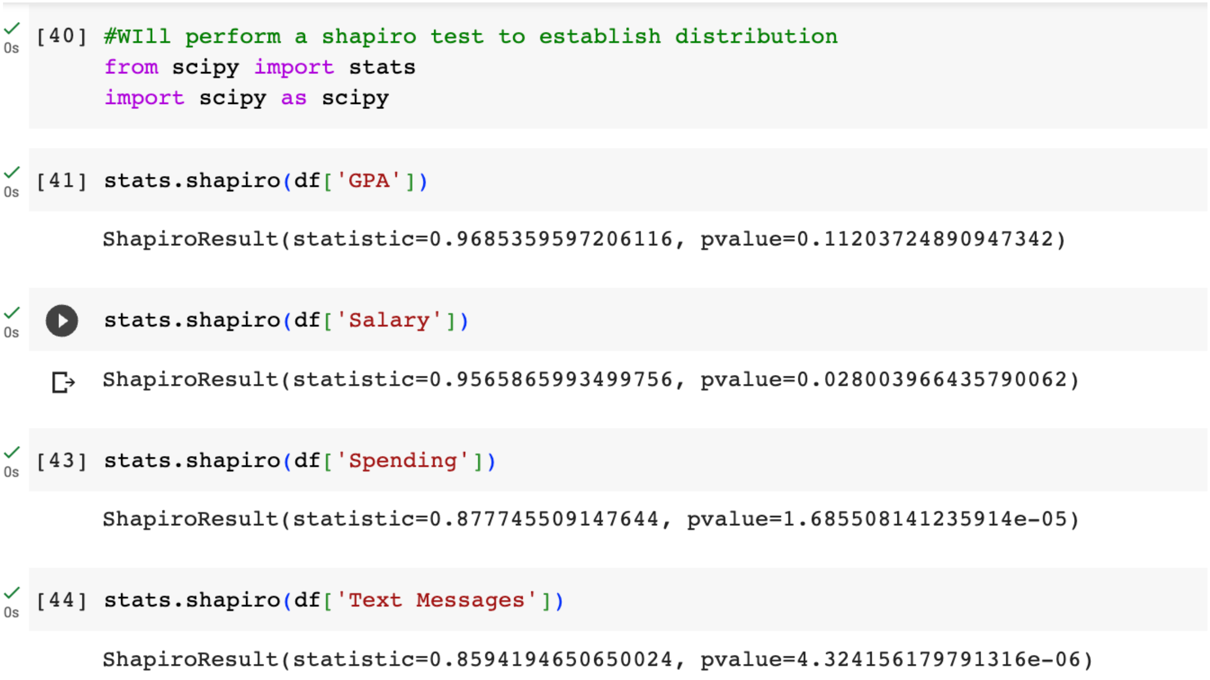
“Spending”



“Text Messages”



Shapiro Test Analysis:



```
[40] #WIll perform a shapiro test to establish distribution
     from scipy import stats
     import scipy as scipy
```

```
[41] stats.shapiro(df['GPA'])
```

```
ShapiroResult(statistic=0.9685359597206116, pvalue=0.11203724890947342)
```

```
stats.shapiro(df['Salary'])
```

```
ShapiroResult(statistic=0.9565865993499756, pvalue=0.028003966435790062)
```

```
[43] stats.shapiro(df['Spending'])
```

```
ShapiroResult(statistic=0.877745509147644, pvalue=1.685508141235914e-05)
```

```
[44] stats.shapiro(df['Text Messages'])
```

```
ShapiroResult(statistic=0.8594194650650024, pvalue=4.324156179791316e-06)
```

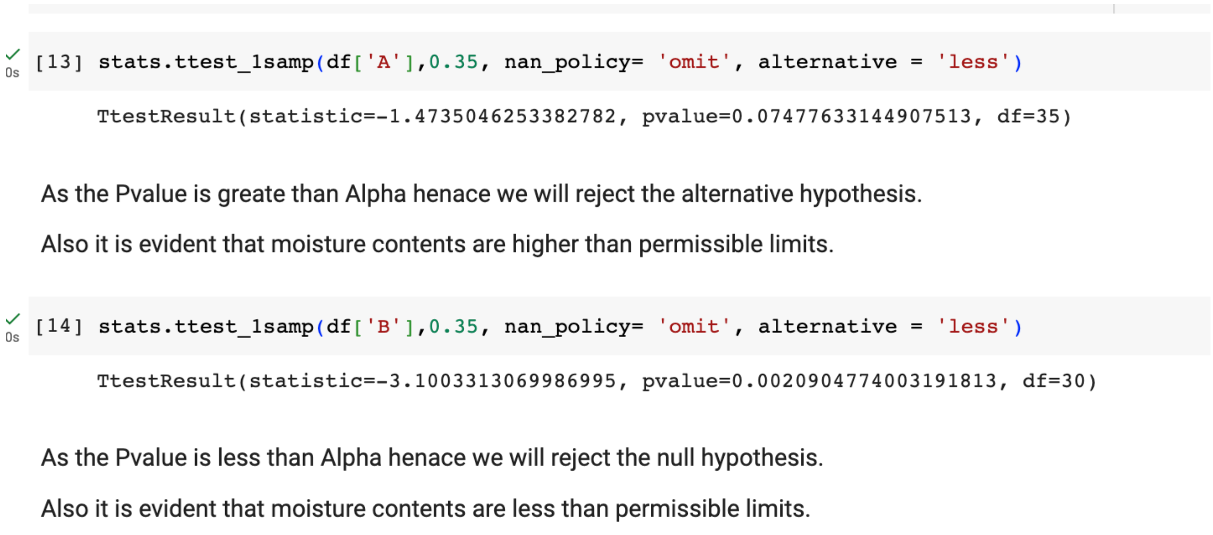
Observation : By these details we confirm that out of the given four data sets ‘GPA’ and ‘Salary’ are normally distributed & ‘Spending’ and ‘Text Messages’ are not.

Project 2

An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amount of moisture the shingles contain when they are packaged. Customers may feel that they have purchased a product lacking in quality if they find moisture and wet shingles inside the packaging. In some cases, excessive moisture can cause the granules attached to the shingles for texture and coloring purposes to fall off the shingles resulting in appearance problems. To monitor the amount of moisture present, the company conducts moisture tests. A shingle is weighed and then dried. The shingle is then reweighed, and based on the amount of moisture taken out of the product, the pounds of moisture per 100 square feet is calculated. The company would like to show that the mean moisture content is less than 0.35 pounds per 100 square feet.
The file includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

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 null hypothesis states that the moisture content of sample A and sample B is greater or equal to the permissible limit (0.35), and the alternative hypothesis states that the moisture content of sample A and sample B is less than permissible limit (0.35).
H0 = >= than 0.35
H1= <= than 0.35



```
[13] stats.ttest_1samp(df['A'],0.35, nan_policy= 'omit', alternative = 'less')

     TtestResult(statistic=-1.4735046253382782, pvalue=0.07477633144907513, df=35)
```

As the Pvalue is greate than Alpha henace we will reject the alternative hypothesis.

Also it is evident that moisture contents are higher than permissible limits.

```
[14] stats.ttest_1samp(df['B'],0.35, nan_policy= 'omit', alternative = 'less')

     TtestResult(statistic=-3.1003313069986995, pvalue=0.0020904774003191813, df=30)
```

As the Pvalue is less than Alpha henace we will reject the null hypothesis.

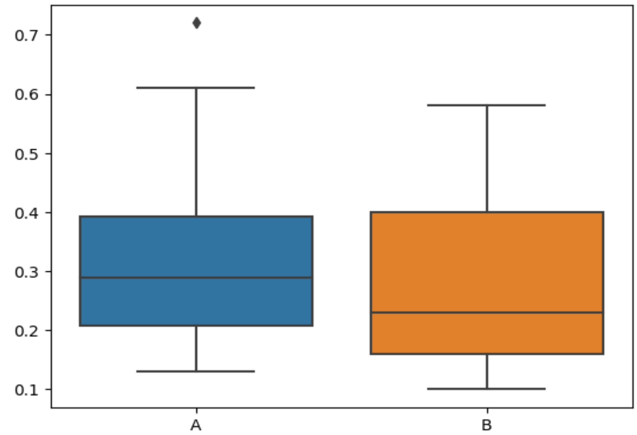
Also it is evident that moisture contents are less than permissible limits.

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?

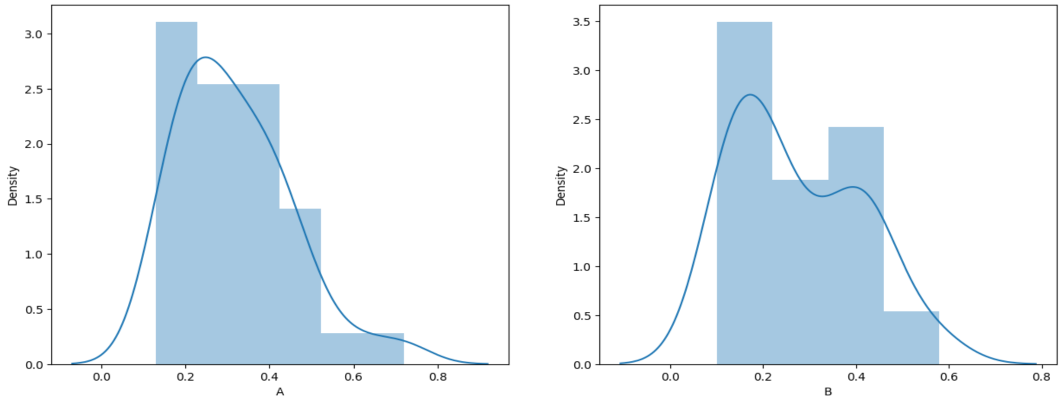
We can frame the Hypotheses as:

Null Hypotheses(H0): μa = μb

Alternative Hypotheses (Ha): μa -= μb

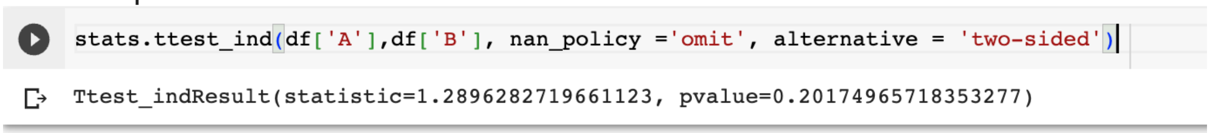


As per box plot the mean of both A & B samples looks different



Data looks like equally distributed

Two sample statistical test:



```
stats.ttest_ind(df['A'],df['B'], nan_policy ='omit', alternative = 'two-sided')
```

```
Ttest_indResult(statistic=1.2896282719661123, pvalue=0.20174965718353277)
```

Here p-value is greater than level of significance so we have to reject the alternative hypothesis in favour of null hypothesis.

We can conclude that mean for shingles-A and singles-B are the same.

Problem 3A:

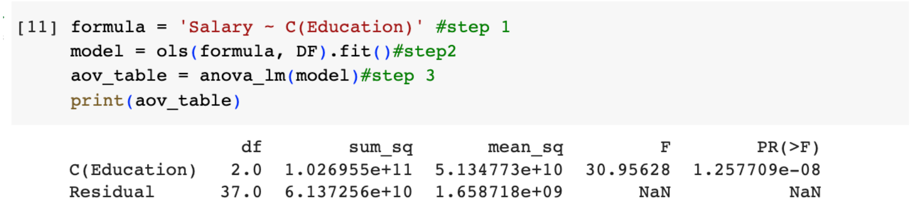
Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals are collected and each person's educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor's, 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.

Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.

1. State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.

Null & Alternate Hypothesis for Education:
H0: Mean of "salary" variable is equal to the mean of "Education" variable.
Ha: at least one of the Mean of "Salary" variable is not equal to the mean of "Education" variable.
***Null & Alternate Hypothesis of Occupation: ***
H0: Mean of "salary" variable is equal to the mean of "Occupation" variable.
Ha: at least one of the Mean of "Salary" variable is not equal to the mean of "Occupation" variable.

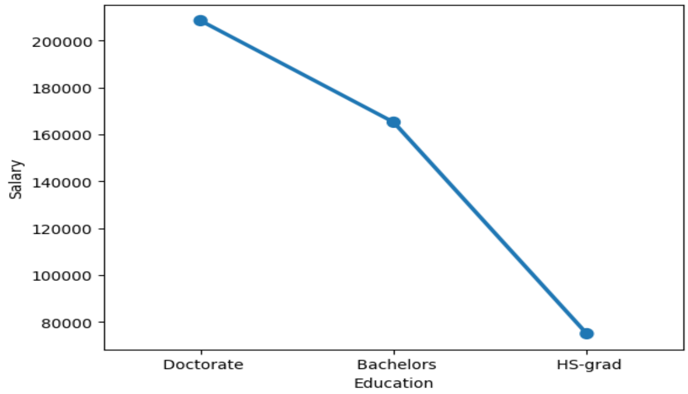
2. Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results.
   Anova Table:



```
[11] formula = 'Salary ~ C(Education)' #step 1
     model = ols(formula, DF).fit()#step2
     aov_table = anova_lm(model)#step 3
     print(aov_table)
```

```
                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
```

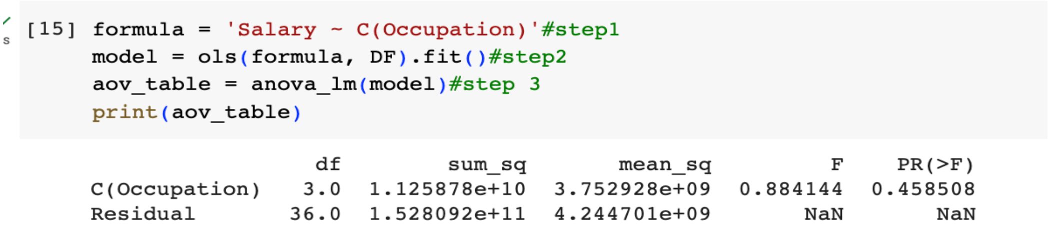
Point plot:



*As the P-Value is less than Alpha hence we reject the null hypothesis. **

3. 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.

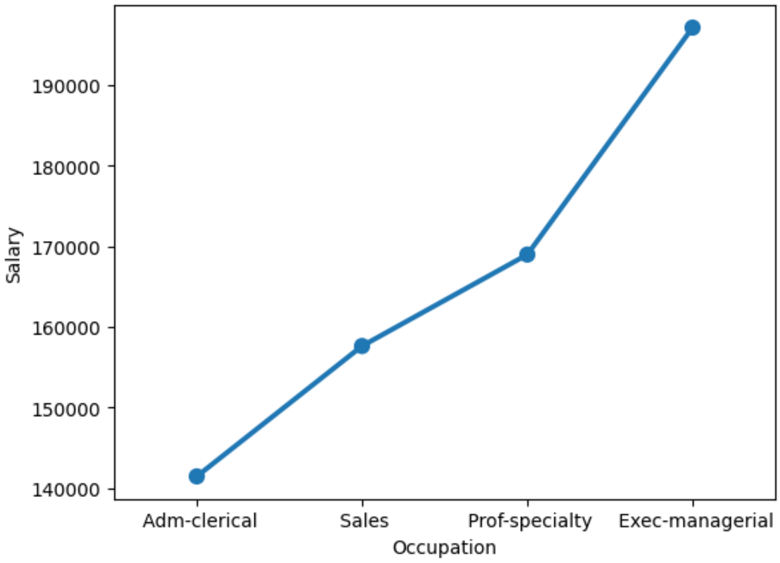
AnovA:



```
[15] formula = 'Salary ~ C(Occupation)'#step1
     model = ols(formula, DF).fit()#step2
     aov_table = anova_lm(model)#step 3
     print(aov_table)
```

```
                  df        sum_sq       mean_sq         F    PR(>F)
C(Occupation)    3.0  1.125878e+10  3.752928e+09  0.884144  0.458508
Residual        36.0  1.528092e+11  4.244701e+09       NaN       NaN
```

Point Plot:



**As the P-Value is greater than Alpha, hence we failed to reject null hypothesis **

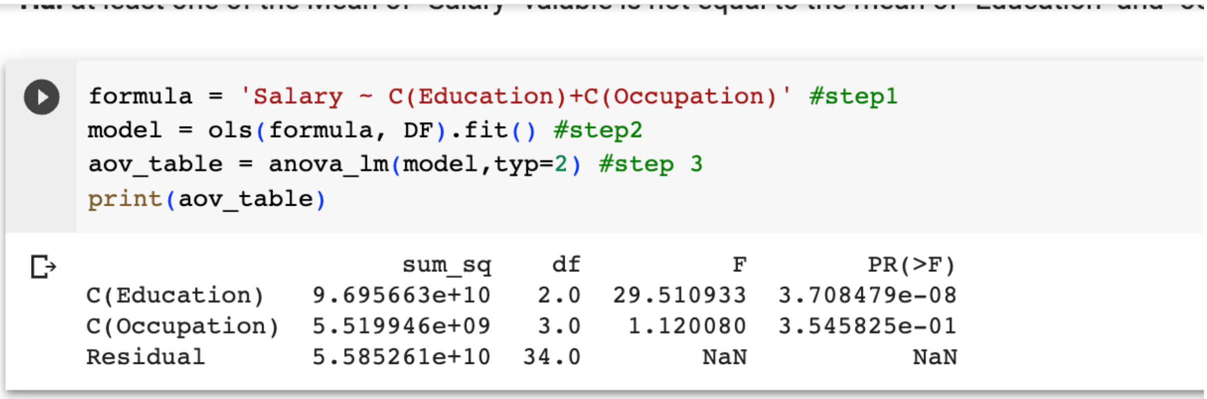
4. If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result. **(Non-Graded)**

5. 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. How will you interpret this result?

   Null & Alternate Hypothesis for Education & Occupation:

   H0: Mean of "salary" variable is equal to the mean of "Education" & Occupation variables.

   Ha: at least one of the Mean of "Salary" variable is not equal to the mean of "Education" and "occupation variables.



Ha: at least one of the Mean of "Salary" variable is not equal to the mean of "Education" and "o

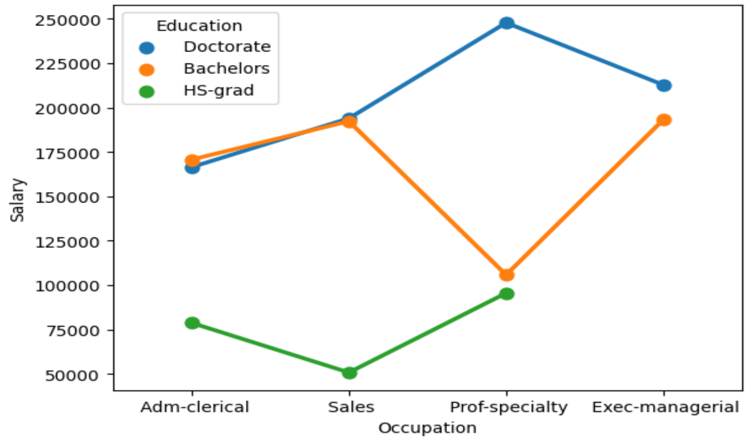
```
formula = 'Salary ~ C(Education)+C(Occupation)' #step1
model = ols(formula, DF).fit() #step2
aov_table = anova_lm(model,typ=2) #step 3
print(aov_table)
```

```
                     sum_sq    df          F        PR(>F)
C(Education)   9.695663e+10   2.0  29.510933  3.708479e-08
C(Occupation)  5.519946e+09   3.0   1.120080  3.545825e-01
Residual       5.585261e+10  34.0        NaN           NaN
```

Interaction between the two categorical variables:

H0: There is no interaction in the means of both the category.

Ha: There is at least one of the group of a category is interaction.



- As seen from the plot above, there seems to interaction between Doctorate and Bachelors on the occupation of ADM-clerical and Sales.
- Also, there is some kind of interaction between Bachelors and HS-grad on the Occupation of Prof-specialty.
- Very minimum to no interaction between Doctorate and HS-grad on the occupation.
- The above indicates that a Doctorate grad may not be highly preferred for a job role and might be over-QUALIFIED which results at par or not significantly higher wage to that of a Bachelor’s degree holder.

- For the variable Education, as the P value is less than significance level which means that Education has a significant impact on the mean Salary. Therefore, Null Hypothesis are rejected.
- Variable Occupation, As the P value is higher than SL of 0.5 hence we failed to reject Null Hypothesis.
- About the interaction, P-Value of Variables “Occupation” + “Education” is less than 0.5 indicating that there is some statistical evidence about the interaction between the 2.

6. Explain the business implications of performing ANOVA for this particular case study.

   Based on the Anova test below are the observations:
   - Salary in significantly dependent on the level of education as compared to the job role.
   - Conclusion of statistics about the interaction between Education and Occupation on Salary, it is safe to say despite occupation less significance there is still some impact of the job role on salary.
   - We also noticed that few bachelors are on higher salary in comparison to doctorate. This may be the result of wrong data provided.