Statistics module assignment

PART A :

Task 1 : Conditional Probability

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)}, \quad P(A) > 0$$

P(S= true | G=female,C=1)
91 / 887 were females in 1st class who survived
94 / 887 were females in 1st class
$\frac{91}{887}$ = 0.102 $\frac{94}{887}$ = 0.105
$P(S\ =\ true \mid G\ =\ female,\ C\ =\ 1)\ \ =\frac{0.102}{0.105}=\ \ 0.971$ = 97.1%

91/887 / 313/887 = 0.103/353 = 0.2917
P(S= true | G=female,C=2)
76 out of 887 are females who were in 2nd class
70 out of 887 are females who were in 2nd class and survived
$\frac{70}{887}$ = 0.78 $\frac{76}{887}$ = 0.86
$P(S\ =\ true \mid G\ =\ female,\ C\ =\ 2)\ \ =\frac{0.078}{0.086}=\ \ 0.906$ = 90.6%

P(S= true | G=female,C=3)
72 / 887 female in 3nd class and survived
144 / 887 female in 3rd class
$\frac{72}{887}$ = 0.081 $\frac{144}{887}$ = 0.162
$P(S\ =\ true \mid G\ =\ female,\ C\ =\ 3)\ \ =\frac{0.081}{0.162}=\ \ 0.5$ = 50%

P(S= true | G=male,C=1)
45 / 887 are males who are in 1st class and survived
122 / 887 are males who were in 1st class
$\frac{45}{887}$ = 0.050 $\frac{122}{887}$ = 0.137
$P(S\ =\ true \mid G\ =\ male,\ C\ =\ 1)\ \ =\frac{0.050}{0.137}=\ \ 0.364$ = 36.4%

P(S= true | G=male,C=2)
108 / 887 are males who were in 2nd class
17 / 887 were males in 2nd class who survived
$\frac{17}{887}$ = 0.019 $\frac{108}{887}$ = 0.121

$$P(S = true \,|\, G = male,\ C = 2) = \frac{0.019}{0.121} = 0.157 = 16\%$$

P(S= true | G=male,C=3)

47 / 887 were males in 3rd class who survived

343 / 887 were males in 3rd class

$\frac{47}{887}$ = 0.052 $\frac{343}{887}$ = 0.386

$$P(S = true \,|\, G = male,\ C = 3) = \frac{0.052}{0.386} = 0.134 = 13.4\%$$

<u>Task b :</u>

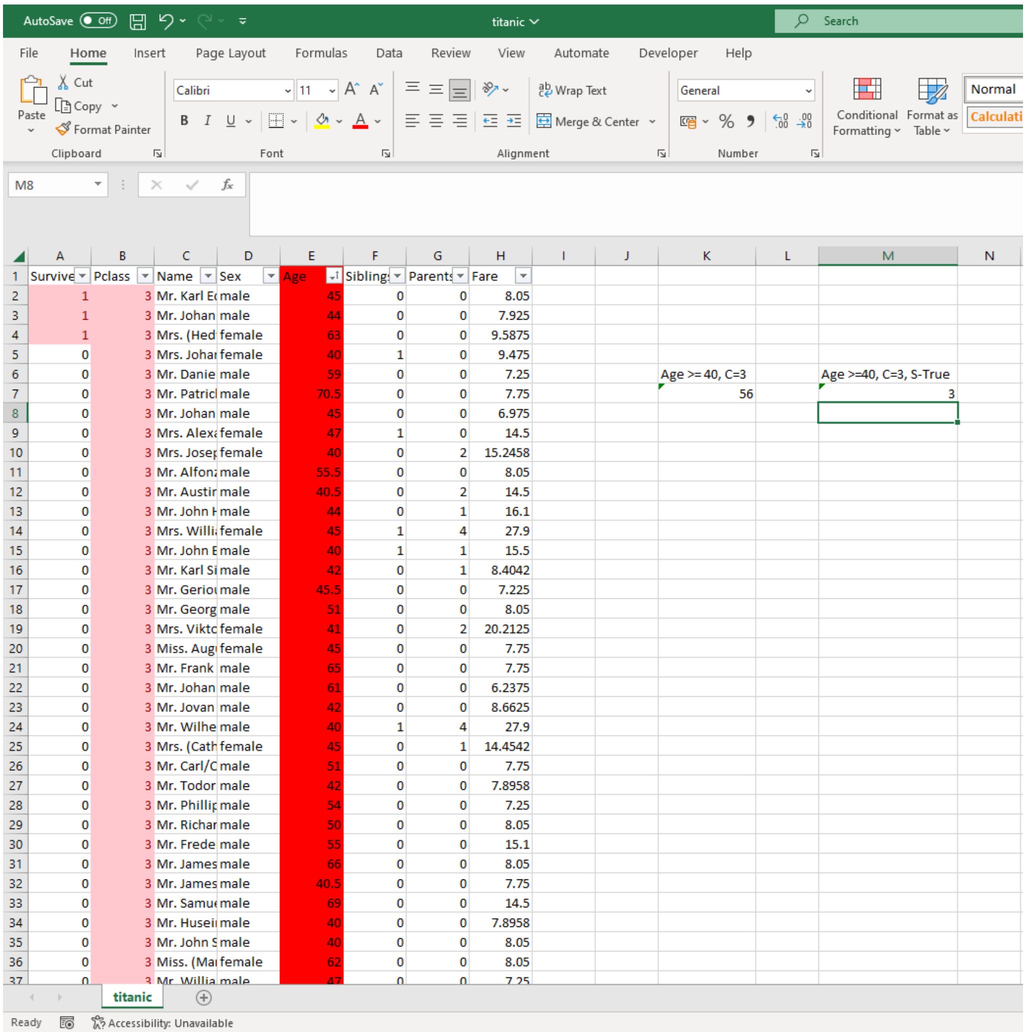
P(S = True | Age >= 40, C = 3)

56 / 887 are 3rd class and have an age of 40 or over

3 / 887 are 3rd class, have an age of 40 or over and survived

$\frac{56}{887}$ = 0.063 $\frac{3}{887}$ = 0.003

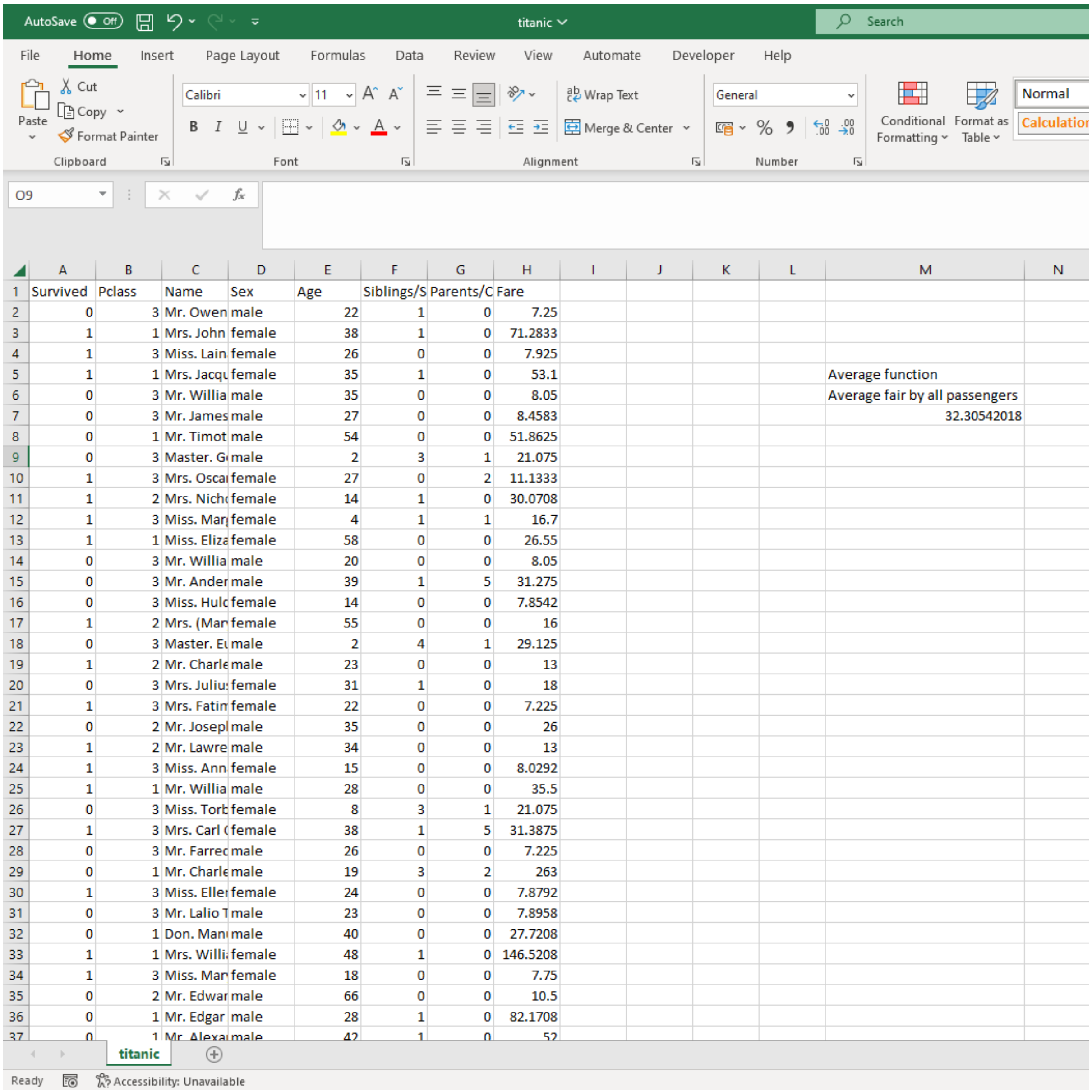
$$P(S = true \,|\, Age >= 40,\ C = 3) = \frac{0.003}{0.063} = 0.047 = 5\%$$



	Survive	Pclass	Name	Sex	Age	Sibling	Parents	Fare
2	1	3	Mr. Karl E	male	45	0	0	8.05
3	1	3	Mr. Johan	male	44	0	0	7.925
4	1	3	Mrs. (Hed	female	63	0	0	9.5875
5	0	3	Mrs. Johar	female	40	1	0	9.475
6	0	3	Mr. Danie	male	59	0	0	7.25
7	0	3	Mr. Patricl	male	70.5	0	0	7.75
8	0	3	Mr. Johan	male	45	0	0	6.975
9	0	3	Mrs. Alexa	female	47	1	0	14.5
10	0	3	Mrs. Josep	female	40	0	2	15.2458
11	0	3	Mr. Alfonz	male	55.5	0	0	8.05
12	0	3	Mr. Austir	male	40.5	0	2	14.5
13	0	3	Mr. John H	male	44	0	1	16.1
14	0	3	Mrs. Willi	female	45	1	4	27.9
15	0	3	Mr. John E	male	40	1	1	15.5
16	0	3	Mr. Karl Si	male	42	0	1	8.4042
17	0	3	Mr. Geriou	male	45.5	0	0	7.225
18	0	3	Mr. Georg	male	51	0	0	8.05
19	0	3	Mrs. Viktc	female	41	0	2	20.2125
20	0	3	Miss. Aug	female	45	0	0	7.75
21	0	3	Mr. Frank	male	65	0	0	7.75
22	0	3	Mr. Johan	male	61	0	0	6.2375
23	0	3	Mr. Jovan	male	42	0	0	8.6625
24	0	3	Mr. Wilhe	male	40	1	4	27.9
25	0	3	Mrs. (Cath	female	45	0	1	14.4542
26	0	3	Mr. Carl/C	male	51	0	0	7.75
27	0	3	Mr. Todor	male	42	0	0	7.8958
28	0	3	Mr. Phillip	male	54	0	0	7.25
29	0	3	Mr. Richar	male	50	0	0	8.05
30	0	3	Mr. Frede	male	55	0	0	15.1
31	0	3	Mr. James	male	66	0	0	8.05
32	0	3	Mr. James	male	40.5	0	0	7.75
33	0	3	Mr. Samue	male	69	0	0	14.5
34	0	3	Mr. Husei	male	40	0	0	7.8958
35	0	3	Mr. John S	male	40	0	0	8.05
36	0	3	Miss. (Mar	female	62	0	0	8.05
37	0	3	Mr. Willia	male	47	0	0	7.25

I used Conditional formatting and highlights in Excel so that i could filter through the data and this would allow me to use =COUNT functions to count the cells that included the data i needed which i could then use later on in my calculations, this was done as counting by hand will present the opportunity for human error which will result in incorrect calculations.

Task c :



	A	B	C	D	E	F	G	H	M
1	Survived	Pclass	Name	Sex	Age	Siblings/S	Parents/C	Fare	
2	0	3	Mr. Owen	male	22	1	0	7.25	
3	1	1	Mrs. John	female	38	1	0	71.2833	
4	1	3	Miss. Lain	female	26	0	0	7.925	
5	1	1	Mrs. Jacqu	female	35	1	0	53.1	Average function
6	0	3	Mr. Willia	male	35	0	0	8.05	Average fair by all passengers
7	0	3	Mr. James	male	27	0	0	8.4583	32.30542018
8	0	1	Mr. Timot	male	54	0	0	51.8625	
9	0	3	Master. G	male	2	3	1	21.075	
10	1	3	Mrs. Osca	female	27	0	2	11.1333	
11	1	2	Mrs. Nich	female	14	1	0	30.0708	
12	1	3	Miss. Mar	female	4	1	1	16.7	
13	1	1	Miss. Eliza	female	58	0	0	26.55	
14	0	3	Mr. Willia	male	20	0	0	8.05	
15	0	3	Mr. Ander	male	39	1	5	31.275	
16	0	3	Miss. Hulc	female	14	0	0	7.8542	
17	1	2	Mrs. (Mar	female	55	0	0	16	
18	0	3	Master. Eu	male	2	4	1	29.125	
19	1	2	Mr. Charle	male	23	0	0	13	
20	0	3	Mrs. Julius	female	31	1	0	18	
21	1	3	Mrs. Fatin	female	22	0	0	7.225	
22	0	2	Mr. Josepl	male	35	0	0	26	
23	1	2	Mr. Lawre	male	34	0	0	13	
24	1	3	Miss. Ann	female	15	0	0	8.0292	
25	1	1	Mr. Willia	male	28	0	0	35.5	
26	0	3	Miss. Torb	female	8	3	1	21.075	
27	1	3	Mrs. Carl (	female	38	1	5	31.3875	
28	0	3	Mr. Farrec	male	26	0	0	7.225	
29	0	1	Mr. Charle	male	19	3	2	263	
30	1	3	Miss. Eller	female	24	0	0	7.8792	
31	0	3	Mr. Lalio T	male	23	0	0	7.8958	
32	0	1	Don. Manı	male	40	0	0	27.7208	
33	1	1	Mrs. Willi	female	48	1	0	146.5208	
34	1	3	Miss. Mar	female	18	0	0	7.75	
35	0	2	Mr. Edwar	male	66	0	0	10.5	
36	0	1	Mr. Edgar	male	28	1	0	82.1708	
37	0	1	Mr. Alexa	male	42	1	0	52	

I simply opened the dataset in Excel and used the =AVG function on all the cells in the "Fare" column as this would efficiently calculate the average fare for me and because this included all passengers i could just highlight all the cells in the column and include them in the calculation. - Average fair paid by all passengers =

*32.30542018 which simplifies to **32.3** (There was no information on currency given so i have not used units)*

Average fare = 32.3

<u>Task 2 : Hypotheses Testing</u>

Survived passengers & class of travel

Chi-Square Calculator

Success! The contingency table below provides the following information: the observed cell totals, (the expected cell totals) and [the chi-square statistic for each cell].

The chi-square statistic, *p*-value and statement of significance appear beneath the table. Blue means you're dealing with dependent variables; red, independent.

Results						
	1st Class	2nd Class	3rd Class			*Row Totals*
Male	45 (43.35) [0.06]	17 (27.73) [4.15]	47 (37.93) [2.17]			109
Female	91 (92.65) [0.03]	70 (59.27) [1.94]	72 (81.07) [1.02]			233
Column Totals	136	87	119			342 (Grand Total)

The chi-square statistic is 9.3711. The *p*-value is .009227. The result is significant at *p* < .05.

```
[1]

from scipy.stats import chi2_contingency

data = [[45, 17, 47], [91, 70, 72]]
stat, p, dof, expected = chi2_contingency(data)

print("The expected values are : ", expected)
print("The chi-squared value is ", stat)
print("The degree of freedom is ", dof)
print("The p value is ", p)

significance_level = 0.05

if p <= significance_level :
    print('Reject HO (There IS an association)')
else:
    print('Accept HO (There IS NOT an association)')
```

```
The expected values are :  [[43.34502924 27.72807018 37.92690058]
 [92.65497076 59.27192982 81.07309942]]
The chi-squred value is  9.371139335460681
The degree of freedom is  2
The p value is  0.00922747651908542
Reject HO (There IS an association)
```

Degree of freedom = (Rows - 1) (Columns - 1)
Degree of freedom = 1 X 2 = 2

D.f =2 / P value = 0.05 = 5.99
9.3711 > 5.99

The Image above tells us that the Chi-square statistic is 9.3711 which is larger than 5.99, this is the number we find on the Chi squared distribution table when looking at a degree of freedom equivalent to 2 and a p value of 0.05. This means that we reject the null hypothesis which is that there is no relationship between the two variables and therefore we accept the alternative hypothesis which is that there is a relationship between the surviving passengers and the class of travel. This outcome suggests that passengers who boarded the Titanic at a higher class(e.g 1st class) were at a much lower risk of death and were subsequently more safe than the passengers in the lower classes, this makes sense as the 1st and 2nd class cabins were located closer to the boat deck(where the lifeboats were located) therefore making access to lifeboats much easier to 1st and 2nd class passengers than 3rd class passengers.

Survived passengers & gender

Chi-Square Calculator

Success! The contingency table below provides the following information: the observed cell totals, (the expected cell totals) and [the chi-square statistic for each cell].

The chi-square statistic, *p*-value and statement of significance appear beneath the table. Blue means you're dealing with dependent variables; red, independent.

Results						
	Male	Female				*Row Totals*
Survived	109 (220.93) [56.71]	233 (121.07) [103.48]				342
Dead	464 (352.07) [35.59]	81 (192.93) [64.94]				545
Column Totals	573	314				887 (Grand Total)

The chi-square statistic is 260.7153. The *p*-value is < .00001. The result is significant at $p < .05$.

```
[6]

from scipy.stats import chi2_contingency

data = [[109, 233], [464, 81]]
stat, p, dof, expected = chi2_contingency(data)

print("The expected values are : ", expected)
print("The chi-squared value is ", stat)
print("The degree of freedom is ", dof)
print("The p value is ", p)

significance_level = 0.05

if p <= significance_level :
    print('Reject HO (There IS an association)')
else:
    print('Accept HO (There IS NOT an association)')
```

```
The expected values are :  [[220.93122886 121.06877114]
 [352.06877114 192.93122886]]
The chi-squared value is  258.39126076789773
The degree of freedom is  1
The p value is  3.847574039733855e-58
Reject HO (There IS an association)
```

Degree of freedom = (Rows - 1) (Columns - 1)
Degree of freedom = 1 X 1 = 1

D.f =1 / P value = 0.05 = 3.84
3.84 < 260.7153

The Image above tells us that the Chi-square statistic is 260.7153 which is larger than 3.84, this is the number we find on the Chi squared distribution table when looking at a degree of freedom equivalent to 1 and a p value of 0.05. This means that we reject the null hypothesis which is that there is no relationship between the passengers that survived and their gender, instead we accept the alternative hypothesis which is that there is a relationship between the passengers that survived and their gender.This outcome suggests that you were more or less likely to survive based on your gender and this could make sense as it is believed that the women and children were given first priority when boarding the lifeboats, when we compile the data, more women survived than men in every class which further backs up this claim.

Task 2 : Visualisation

Create multiple individual graphs

Ready

Run notebook

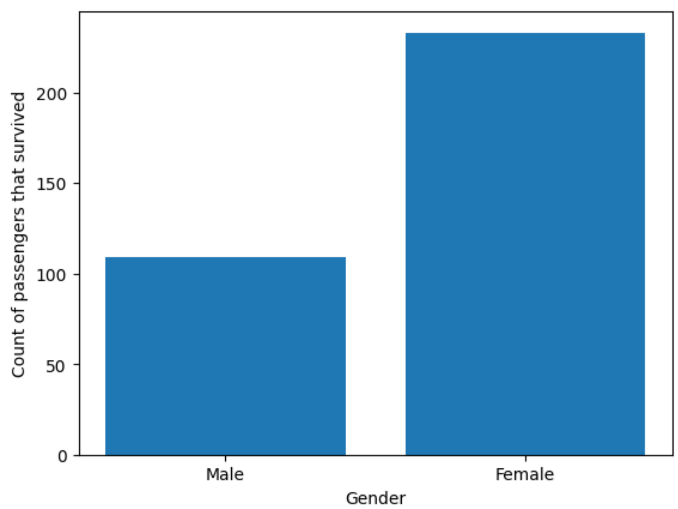
[9]

```
# Below we are importing the libraries needed to create and plot data on the graph
import numpy as np
from matplotlib import pyplot

Category_of_passenger = ('Survived M+F', 'Dead M+F', '1st class', '2nd class', '3rd class') # Names of each bar
y_val = np.arange(len(Category_of_passenger))
Category_values = [342, 544, 216, 184, 487] # Assigning values to each bar

pyplot.bar(y_val, Category_values, align='center') # pyplot.bar signifies a bar chart
pyplot.xticks(y_val, Category_of_passenger)
pyplot.ylabel('Count of passengers') # Naming the y axis
pyplot.xlabel('Category of passenger') # Naming the x axis

pyplot.show() # Outputs the graph we described above
```

Code Text SQL Chart Input

This graph shows all the categories and their corresponding values in one bar graph

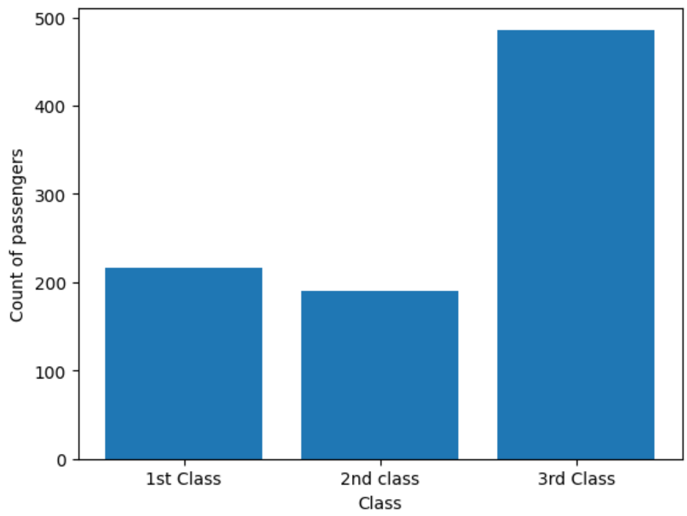
```python
# Below we are importing the libraries needed to create and plot data on the graph
import numpy as np
from matplotlib import pyplot

Category_of_passenger = ('Male', 'Female') # Names of each bar
y_val = np.arange(len(Category_of_passenger))
Category_values = [109, 233] # Assigning values to each bar

pyplot.bar(y_val, Category_values, align='center') # pyplot.bar signifies a bar chart
pyplot.xticks(y_val, Category_of_passenger)
pyplot.ylabel('Count of passengers that survived') # Naming the y axis
pyplot.xlabel('Gender') # Naming the x axis

pyplot.show() # Outputs the graph we described above
```



Code | Text | SQL | Chart | Input

This graph shows the number of males and females that survived

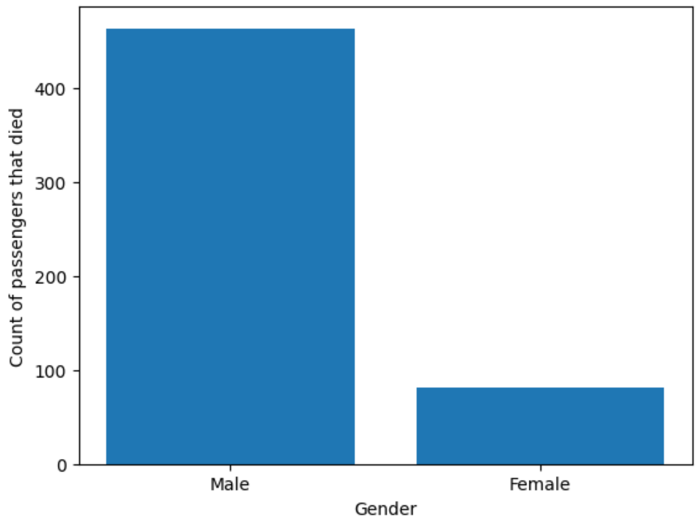
```
# Below we are importing the libraries needed to create and plot data on the graph
import numpy as np
from matplotlib import pyplot

Category_of_passenger = ('1st Class', '2nd class', '3rd Class') # Names of each bar
y_val = np.arange(len(Category_of_passenger))
Category_values = [216, 190, 486] # Assigning values to each bar

pyplot.bar(y_val, Category_values, align='center') # pyplot.bar signifies a bar chart
pyplot.xticks(y_val, Category_of_passenger)
pyplot.ylabel('Count of passengers') # Naming the y axis
pyplot.xlabel('Class') # Naming the x axis

pyplot.show() # Outputs the graph we described above
```



Code Text SQL Chart Input

This graph shows the number of passengers in each class

```python
# Below we are importing the libraries needed to create and plot data on the graph
import numpy as np
from matplotlib import pyplot

Category_of_passenger = ('Male', 'Female') # Names of each bar
y_val = np.arange(len(Category_of_passenger))
Category_values = [463, 81] # Assigning values to each bar

pyplot.bar(y_val, Category_values, align='center') # pyplot.bar signifies a bar chart
pyplot.xticks(y_val, Category_of_passenger)
pyplot.ylabel('Count of passengers that died') # Naming the y axis
pyplot.xlabel('Gender') # Naming the x axis

pyplot.show() # Outputs the graph we described above
```

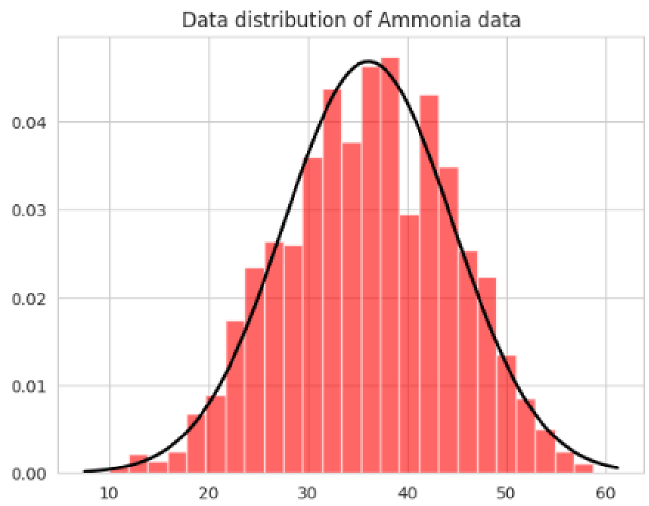


Code Text SQL Chart Input

This graph shows the number of males and females that died

PART B : Data distribution

```
import numpy as np
from scipy.stats import norm
import matplotlib.pyplot as plt
import pandas as pd


# Reading the CSV file called 'ammonia' and assigning it to variable 'data'
data = pd.read_csv('ammonia.csv')

# Plot the histogram to show us an overall trend
plt.hist(data, bins=25, density=True, alpha=0.6, color='red')

# Fitting a normal distribution to the data:
mu, std = norm.fit(data)

# Plot the PDF.
xmin, xmax = plt.xlim()
x = np.linspace(xmin, xmax, 100)
p = norm.pdf(x, mu, std)
plt.plot(x, p, 'k', linewidth=2)
title = "Data distribution of Ammonia data"
plt.title(title)

plt.show()
```



The image above is the coding i have created on Deep note, i have also included the output which shows a histogram of the ammonia data from ‘ammonia.csv’, from the shape of the bars in the histogram i can infer that there is a normal distribution, i then added some code to overlay a bell curve to further make my observation more apparent.

To answer the questions :

- I believe a normal distribution is the most likely distribution of the dataset.
- I have estimated the mean to be 36.1 and the standard deviation to be 8.51 assuming that the data is from a normal distribution.

The probability of the ammonia concentration being greater than 30 and less than 40 mg/L is 44%.

$$\text{Zscore} = \frac{(x - \mu)}{\sigma}$$

> 30 and < 40

$$\frac{40 - 36.094993}{8.515969} = 0.45855111$$

$$\frac{30 - 36.094993}{8.515969} = -0.7157134$$

$$0.45855111 = 0.6736$$

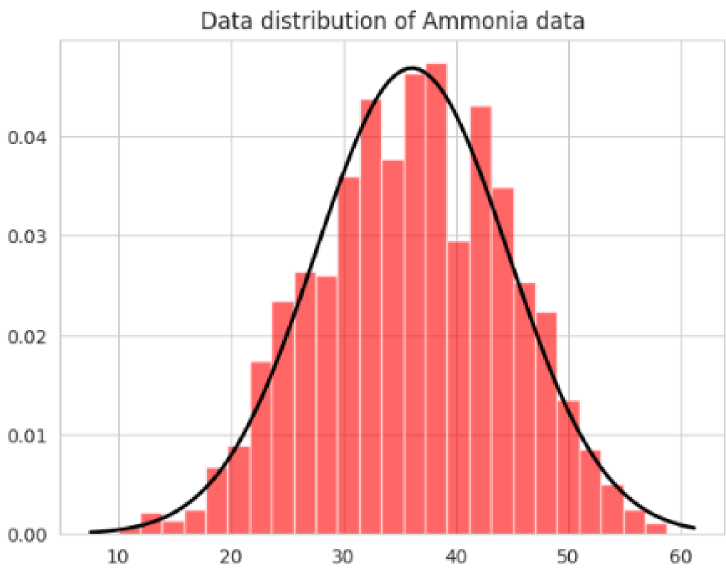
$$-0.7157134 = 0.2358$$

$$0.6736 - 0.2358 = 0.4378$$

$$= 44\%$$

Above is the manual calculation for the probability of the ammonia concentration being greater than 30 and less than 40 mg/L. Below is the python calculation of the probability and this also shows 43.96 % which rounded up is 44%

```
Less_than_40 = norm(loc = 36.094993 , scale = 8.515969).cdf(40)

Greater_than_30 = norm(loc = 36.094993 , scale = 8.515969).cdf(30)


Probability = Less_than_40 - Greater_than_30
print(Probability * 100)

mean = np.mean(data)
sd = np.std(data)
print(sd)
print(mean)

print(Less_than_40)
```




```
43.96375503825954
Ammonia    8.515969
dtype: float64
Ammonia    36.094993
dtype: float64
0.6767217224658089
```

PART C : Combination probability

$$n_r^c = \frac{n!}{r!(n-r)!}$$

3 consonants out of 7
2 Vowels out of 4

$7C_3$ and $4C_2$

$$7C_3 \times 4C_2 = \frac{7!}{4!\,3!} \times \frac{4!}{2!\,2!}$$

$$= \frac{7 \times 6 \times 5}{3 \times 2} \times \frac{4 \times 3}{2 \times 1} = 35 \times 6 = 210$$

$3+2 = 5$
5 words in itself $= 5! = 120$

$210 \times 120 = 25200$

25200 possible combinations.

APPENDIX

Graph showing all categories

```
# Below we are importing the libraries needed to create and plot data on
the graph
import numpy as np
from matplotlib import pyplot

Category_of_passenger = ('Survived M+F', 'Dead M+F', '1st class', '2nd
class', '3rd class') # Names of each bar
y_val = np.arange(len(Category_of_passenger))
Category_values = [342, 544, 216, 184, 487] # Assigning values to each bar
```

```
pyplot.bar(y_val, Category_values, align='center') # pyplot.bar signifies a bar chart
pyplot.xticks(y_val, Category_of_passenger)
pyplot.ylabel('Count of passengers') # Naming the y axis
pyplot.xlabel('Category of passenger') # Naming the x axis

pyplot.show() # Outputs the graph we described above
```

——--

Male and females that survived

```
# Below we are importing the libraries needed to create and plot data on the graph
import numpy as np
from matplotlib import pyplot

Category_of_passenger = ('Male', 'Female') # Names of each bar
y_val = np.arange(len(Category_of_passenger))
Category_values = [109, 233] # Assigning values to each bar

pyplot.bar(y_val, Category_values, align='center') # pyplot.bar signifies a bar chart
pyplot.xticks(y_val, Category_of_passenger)
pyplot.ylabel('Count of passengers that survived') # Naming the y axis
pyplot.xlabel('Gender') # Naming the x axis

pyplot.show() # Outputs the graph we described above
```

——--

Male and females that died

```
# Below we are importing the libraries needed to create and plot data on the graph
import numpy as np
from matplotlib import pyplot

Category_of_passenger = ('Male', 'Female') # Names of each bar
y_val = np.arange(len(Category_of_passenger))
Category_values = [463, 81] # Assigning values to each bar
```

```
pyplot.bar(y_val, Category_values, align='center') # pyplot.bar signifies a
bar chart
pyplot.xticks(y_val, Category_of_passenger)
pyplot.ylabel('Count of passengers that died') # Naming the y axis
pyplot.xlabel('Gender') # Naming the x axis

pyplot.show() # Outputs the graph we described above
```

——---

Count of passengers and what class they were in

```
# Below we are importing the libraries needed to create and plot data on
the graph
import numpy as np
from matplotlib import pyplot

Category_of_passenger = ('1st Class', '2nd class', '3rd Class') # Names of
each bar
y_val = np.arange(len(Category_of_passenger))
Category_values = [216, 190, 486] # Assigning values to each bar

pyplot.bar(y_val, Category_values, align='center') # pyplot.bar signifies a
bar chart
pyplot.xticks(y_val, Category_of_passenger)
pyplot.ylabel('Count of passengers') # Naming the y axis
pyplot.xlabel('Class') # Naming the x axis

pyplot.show() # Outputs the graph we described above
```

——--

Data distribution graph

```
import numpy as np
from scipy.stats import norm
import matplotlib.pyplot as plt
import pandas as pd
```

```
# Reading the CSV file called 'ammonia' and assigning it to variable 'data'
data = pd.read_csv('ammonia.csv')

# Plot the histogram to show us an overall trend
plt.hist(data, bins=25, density=True, alpha=0.6, color='red')

# Fitting a normal distribution to the data:
mu, std = norm.fit(data)

# Plot the PDF.
xmin, xmax = plt.xlim()
x = np.linspace(xmin, xmax, 100)
p = norm.pdf(x, mu, std)
plt.plot(x, p, 'k', linewidth=2)
title = "Data distribution of Ammonia data"
plt.title(title)

plt.show()

#Less_than_40 = norm(loc = 36.094993 , scale = 8.515969).cdf(40)

#Greater_than_30 = norm(loc = 36.094993 , scale = 8.515969).cdf(30)


#Probability = Less_than_40 - Greater_than_30
#print(Probability * 100)

#mean = np.mean(data)
#sd = np.std(data)
#print(sd)
#print(mean)

#print(Less_than_40)
```

——--

Chi squared test #1

```
from scipy.stats import chi2_contingency

data = [[45, 17, 47], [91, 70, 72]]
```

```python
from scipy.stats import chi2_contingency

data = [[45, 17, 47], [91, 70, 72]]
stat, p, dof, expected = chi2_contingency(data)

print("The expected values are : ", expected)
print("The chi-squared value is ", stat)
print("The degree of freedom is ", dof)
print("The p value is ", p)

significance_level = 0.05

if p <= significance_level :
    print('Reject HO (There IS an association)')
else:
    print('Accept HO (There IS NOT an association)')
```

Chi squared test #2

```python
from scipy.stats import chi2_contingency

data = [[109, 233], [464, 81]]
stat, p, dof, expected = chi2_contingency(data)

print("The expected values are : ", expected)
print("The chi-squared value is ", stat)
print("The degree of freedom is ", dof)
print("The p value is ", p)

significance_level = 0.05

if p <= significance_level :
    print('Reject HO (There IS an association)')
else:
    print('Accept HO (There IS NOT an association)')
```