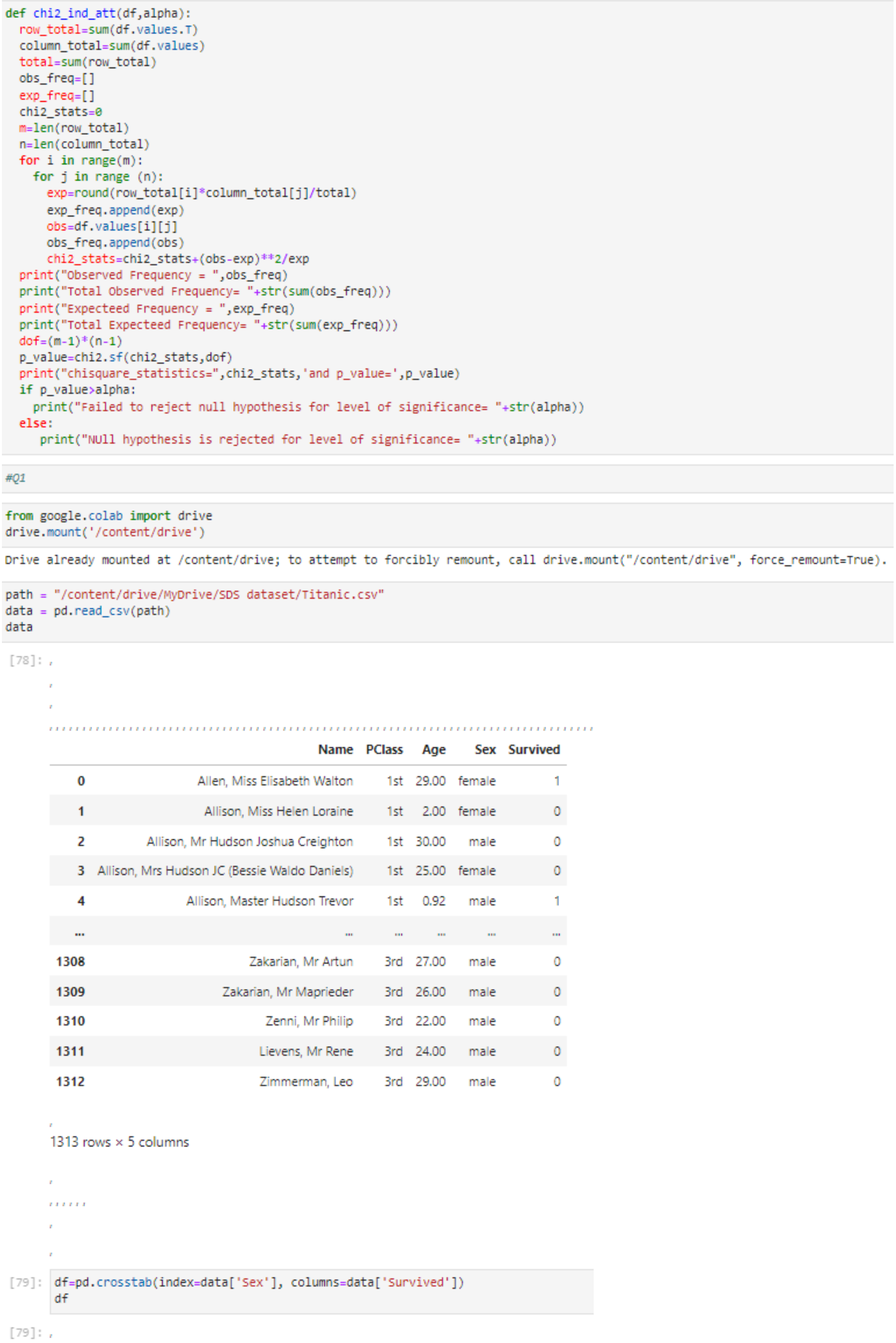
Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)
B.Tech. Sem: III Subject: Statistics for Data Science
Experiment 8

Name: Rishabh Singhvi **SAP ID: 60009210206**

Date:	Experiment Title: Chisquare Test using Python
Aim	To study test of independence of attributes and goodness of fit.
Software	Google Colab
Theory	**Question 1:** For a give Titanic dataset, can it be concluded that gender and survival of passengers are related to each other?



```
def chi2_ind_att(df,alpha):
  row_total=sum(df.values.T)
  column_total=sum(df.values)
  total=sum(row_total)
  obs_freq=[]
  exp_freq=[]
  chi2_stats=0
  m=len(row_total)
  n=len(column_total)
  for i in range(m):
    for j in range (n):
      exp=round(row_total[i]*column_total[j]/total)
      exp_freq.append(exp)
      obs=df.values[i][j]
      obs_freq.append(obs)
      chi2_stats=chi2_stats+(obs-exp)**2/exp
  print("Observed Frequency = ",obs_freq)
  print("Total Observed Frequency= "+str(sum(obs_freq)))
  print("Expecteed Frequency = ",exp_freq)
  print("Total Expecteed Frequency= "+str(sum(exp_freq)))
  dof=(m-1)*(n-1)
  p_value=chi2.sf(chi2_stats,dof)
  print("chisquare_statistics=",chi2_stats,'and p_value=',p_value)
  if p_value>alpha:
    print("Failed to reject null hypothesis for level of significance= "+str(alpha))
  else:
     print("NUll hypothesis is rejected for level of significance= "+str(alpha))
```

```
#Q1
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
path = "/content/drive/MyDrive/SDS dataset/Titanic.csv"
data = pd.read_csv(path)
data
```

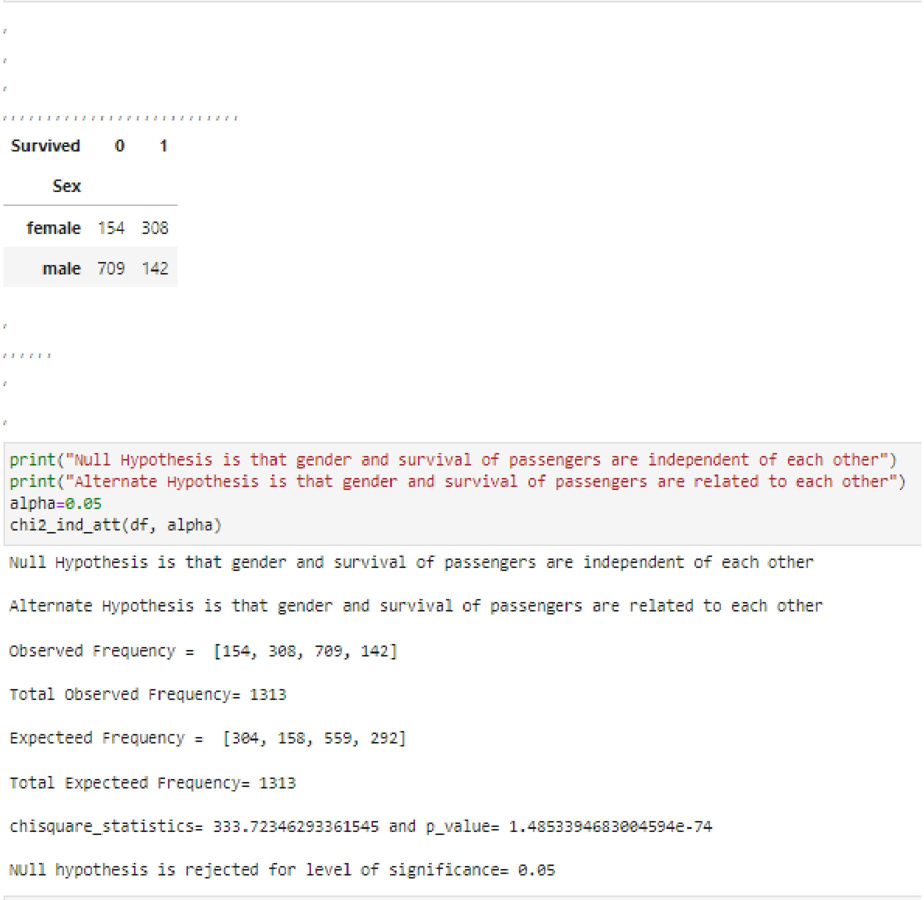
[78]: ,

	Name	PClass	Age	Sex	Survived
0	Allen, Miss Elisabeth Walton	1st	29.00	female	1
1	Allison, Miss Helen Loraine	1st	2.00	female	0
2	Allison, Mr Hudson Joshua Creighton	1st	30.00	male	0
3	Allison, Mrs Hudson JC (Bessie Waldo Daniels)	1st	25.00	female	0
4	Allison, Master Hudson Trevor	1st	0.92	male	1
...	...	...	...	...	...
1308	Zakarian, Mr Artun	3rd	27.00	male	0
1309	Zakarian, Mr Maprieder	3rd	26.00	male	0
1310	Zenni, Mr Philip	3rd	22.00	male	0
1311	Lievens, Mr Rene	3rd	24.00	male	0
1312	Zimmerman, Leo	3rd	29.00	male	0

1313 rows × 5 columns

[79]:
```
df=pd.crosstab(index=data['Sex'], columns=data['Survived'])
df
```

[79]: ,



,
,
,
,,,,,,,,,,,,,,,,,,,,,,,,,,

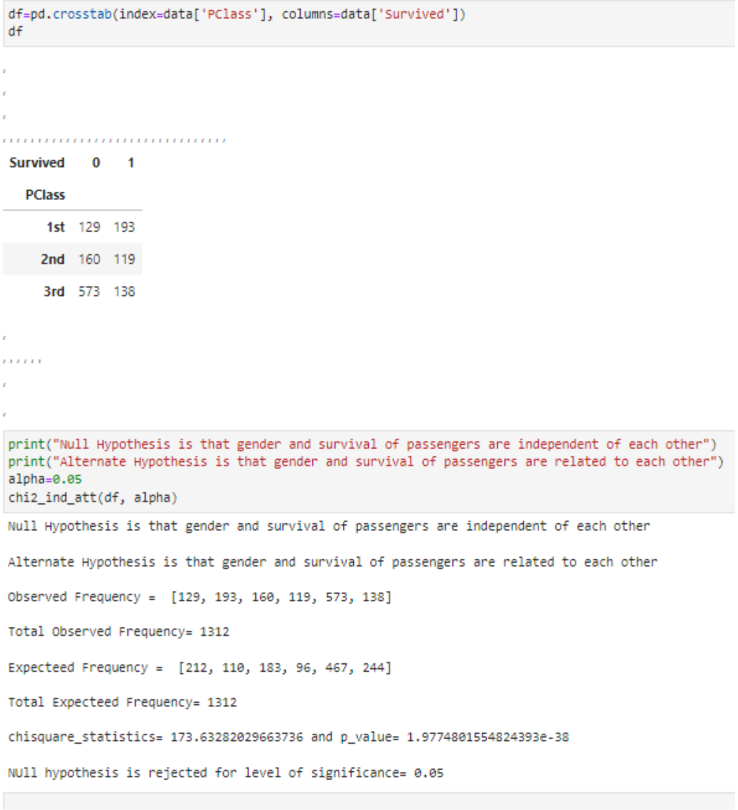
Survived	0	1
Sex		
female	154	308
male	709	142

,
,,,,,,
,
,

```
print("Null Hypothesis is that gender and survival of passengers are independent of each other")
print("Alternate Hypothesis is that gender and survival of passengers are related to each other")
alpha=0.05
chi2_ind_att(df, alpha)
```

```
Null Hypothesis is that gender and survival of passengers are independent of each other

Alternate Hypothesis is that gender and survival of passengers are related to each other

Observed Frequency =  [154, 308, 709, 142]

Total Observed Frequency= 1313

Expecteed Frequency =  [304, 158, 559, 292]

Total Expecteed Frequency= 1313

chisquare_statistics= 333.72346293361545 and p_value= 1.4853394683004594e-74

NUll hypothesis is rejected for level of significance= 0.05
```

Question 2: For a give Titanic dataset, can it be concluded that class and survival of passengers are related to each other?



```
df=pd.crosstab(index=data['PClass'], columns=data['Survived'])
df
```

,
,
,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,

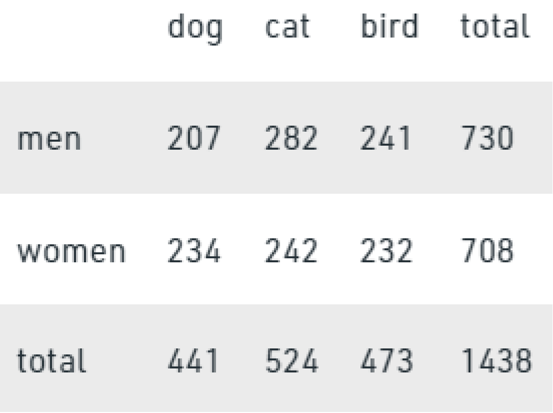
Survived	0	1
PClass		
1st	129	193
2nd	160	119
3rd	573	138

,
,,,,,,
,
,

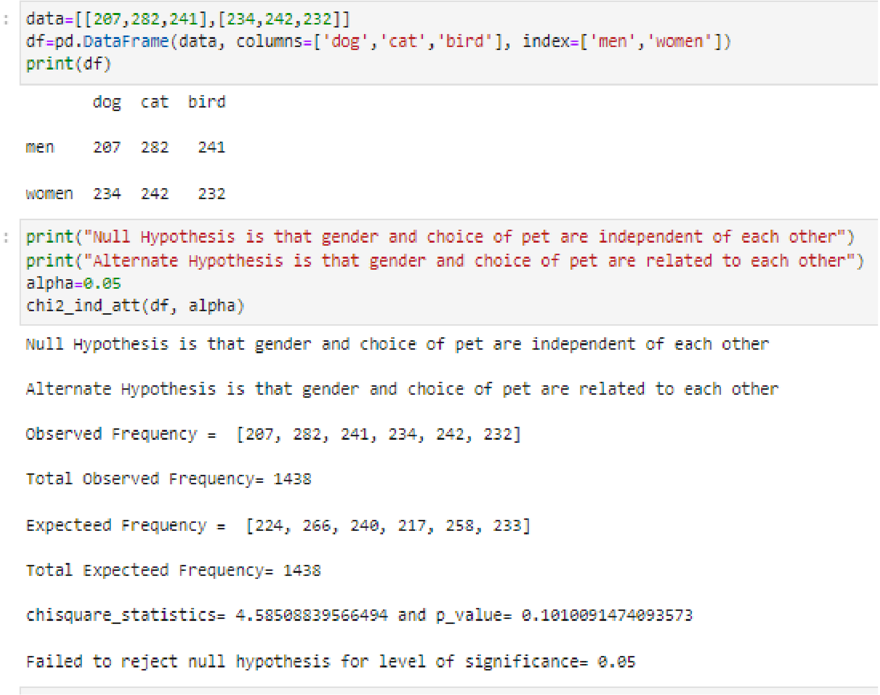
```
print("Null Hypothesis is that gender and survival of passengers are independent of each other")
print("Alternate Hypothesis is that gender and survival of passengers are related to each other")
alpha=0.05
chi2_ind_att(df, alpha)
```

```
Null Hypothesis is that gender and survival of passengers are independent of each other

Alternate Hypothesis is that gender and survival of passengers are related to each other

Observed Frequency =  [129, 193, 160, 119, 573, 138]

Total Observed Frequency= 1312

Expecteed Frequency =  [212, 110, 183, 96, 467, 244]

Total Expecteed Frequency= 1312

chisquare_statistics= 173.63282029663736 and p_value= 1.9774801554824393e-38

NUll hypothesis is rejected for level of significance= 0.05
```

Question 3: A table shows the number of men and women buying different types of pets. Can it be concluded that gender and choice of pet are related to each other?



	dog	cat	bird	total
men	207	282	241	730
women	234	242	232	708
total	441	524	473	1438



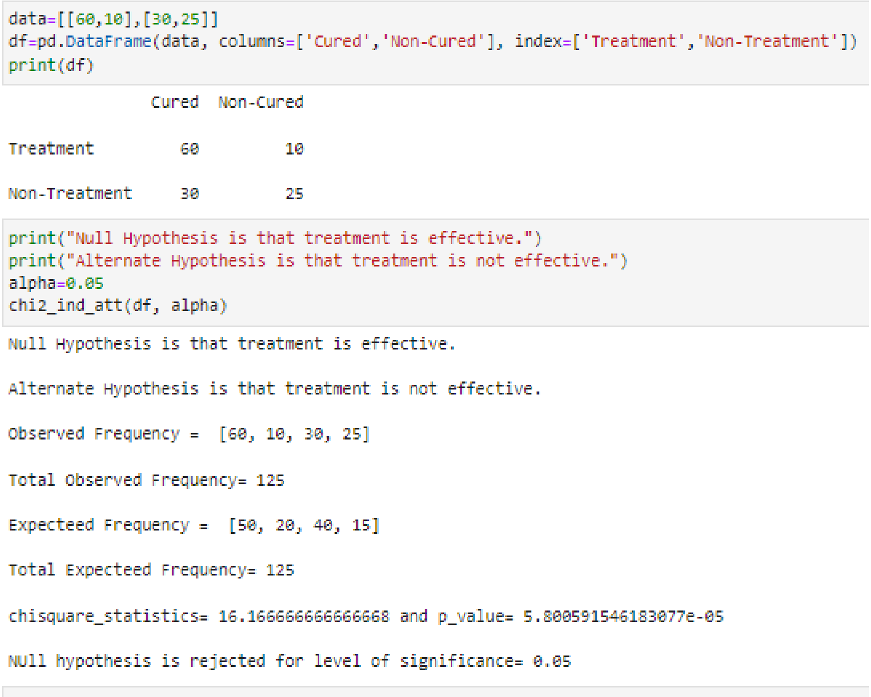
```
data=[[207,282,241],[234,242,232]]
df=pd.DataFrame(data, columns=['dog','cat','bird'], index=['men','women'])
print(df)
```

```
       dog  cat  bird

men    207  282   241

women  234  242   232
```

```
print("Null Hypothesis is that gender and choice of pet are independent of each other")
print("Alternate Hypothesis is that gender and choice of pet are related to each other")
alpha=0.05
chi2_ind_att(df, alpha)
```

```
Null Hypothesis is that gender and choice of pet are independent of each other

Alternate Hypothesis is that gender and choice of pet are related to each other

Observed Frequency =  [207, 282, 241, 234, 242, 232]

Total Observed Frequency= 1438

Expecteed Frequency =  [224, 266, 240, 217, 258, 233]

Total Expecteed Frequency= 1438

chisquare_statistics= 4.58508839566494 and p_value= 0.1010091474093573

Failed to reject null hypothesis for level of significance= 0.05
```

Question 4: For the given drug data, can we conclude that treatment is effective?



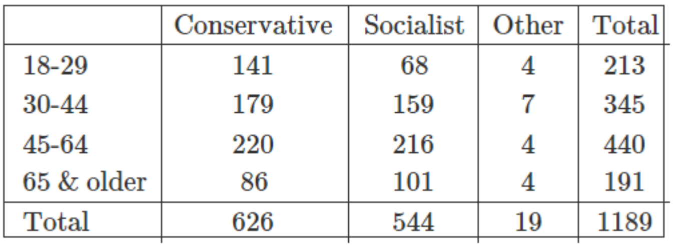
```
data=[[60,10],[30,25]]
df=pd.DataFrame(data, columns=['Cured','Non-Cured'], index=['Treatment','Non-Treatment'])
print(df)
```

```
               Cured  Non-Cured

Treatment         60         10

Non-Treatment     30         25
```

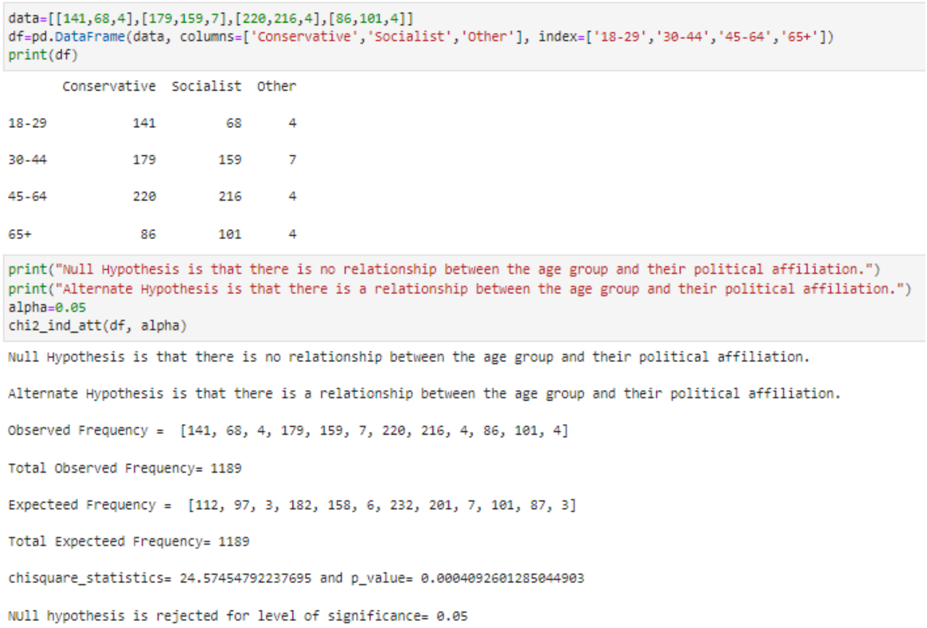
```
print("Null Hypothesis is that treatment is effective.")
print("Alternate Hypothesis is that treatment is not effective.")
alpha=0.05
chi2_ind_att(df, alpha)
```

```
Null Hypothesis is that treatment is effective.

Alternate Hypothesis is that treatment is not effective.

Observed Frequency =  [60, 10, 30, 25]

Total Observed Frequency= 125

Expecteed Frequency =  [50, 20, 40, 15]

Total Expecteed Frequency= 125

chisquare_statistics= 16.166666666666668 and p_value= 5.800591546183077e-05

NUll hypothesis is rejected for level of significance= 0.05
```

Question 5: The table below is an exit poll which displays the joint responses to 2 categorical variables: people in categories from 18–29, 30–44, 45–64 and >65 years, and their political affiliation, which is "Conservative", "Socialist" and "Other". Create data corresponding to this information. Is there any evidence of a relationship between the age group and their political affiliation, at 5% significant level?



	Conservative	Socialist	Other	Total
18-29	141	68	4	213
30-44	179	159	7	345
45-64	220	216	4	440
65 & older	86	101	4	191
Total	626	544	19	1189



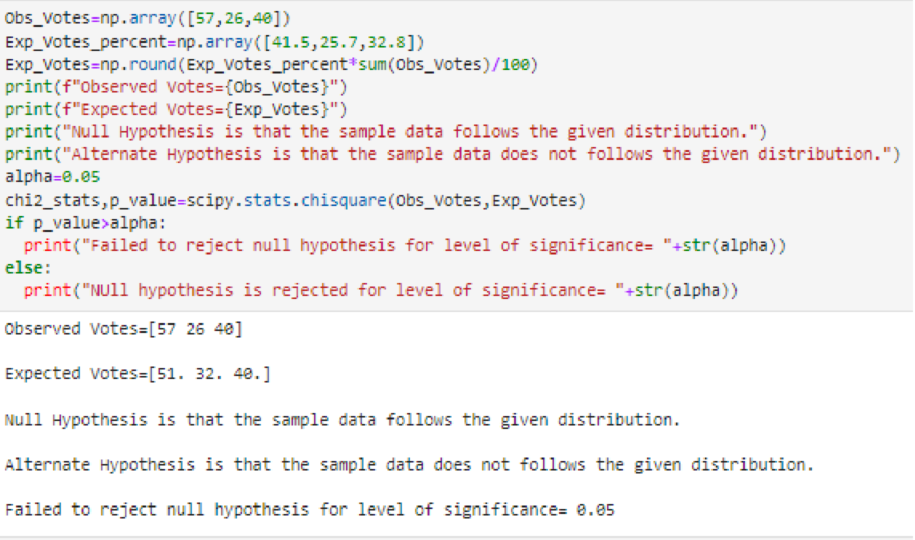
```
data=[[141,68,4],[179,159,7],[220,216,4],[86,101,4]]
df=pd.DataFrame(data, columns=['Conservative','Socialist','Other'], index=['18-29','30-44','45-64','65+'])
print(df)
```

```
       Conservative  Socialist  Other
18-29           141         68      4
30-44           179        159      7
45-64           220        216      4
65+              86        101      4
```

```
print("Null Hypothesis is that there is no relationship between the age group and their political affiliation.")
print("Alternate Hypothesis is that there is a relationship between the age group and their political affiliation.")
alpha=0.05
chi2_ind_att(df, alpha)
```

```
Null Hypothesis is that there is no relationship between the age group and their political affiliation.

Alternate Hypothesis is that there is a relationship between the age group and their political affiliation.

Observed Frequency =  [141, 68, 4, 179, 159, 7, 220, 216, 4, 86, 101, 4]

Total Observed Frequency= 1189

Expecteed Frequency =  [112, 97, 3, 182, 158, 6, 232, 201, 7, 101, 87, 3]

Total Expecteed Frequency= 1189

chisquare_statistics= 24.57454792237695 and p_value= 0.0004092601285044903

NUll hypothesis is rejected for level of significance= 0.05
```

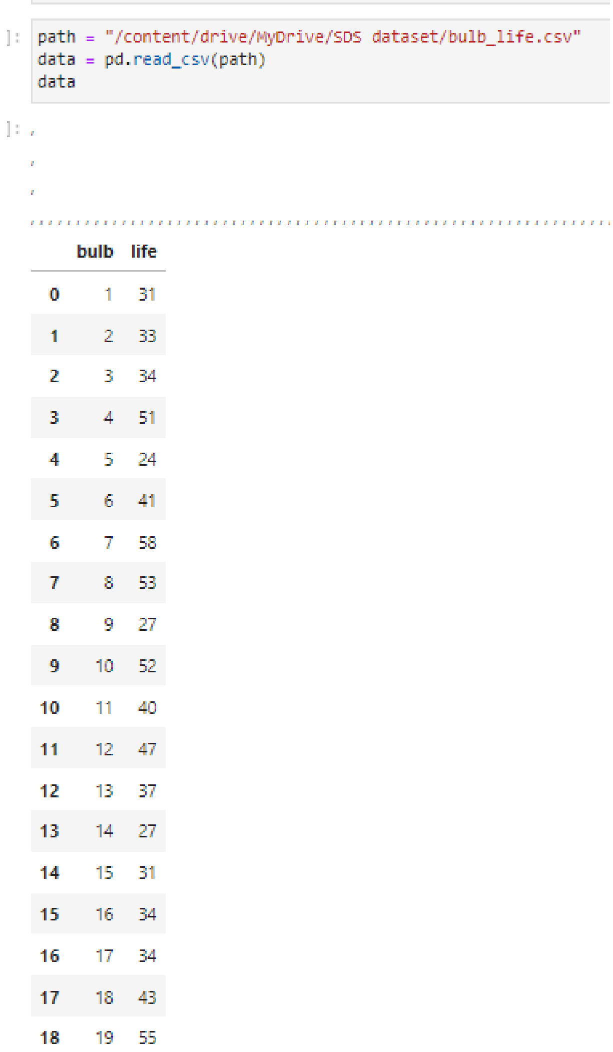
Question 6: A researcher takes a random sample and pick 123 students about their party affiliation. Out of them 57 vote for party A, 26 vote for party B and 40 for Others. Generally, 41.5% of people vote for the party A, 25.7% for the party B and the remaining 32.8% as Others. Test the hypothesis that sample data follows given distribution.



```
Obs_Votes=np.array([57,26,40])
Exp_Votes_percent=np.array([41.5,25.7,32.8])
Exp_Votes=np.round(Exp_Votes_percent*sum(Obs_Votes)/100)
print(f"Observed Votes={Obs_Votes}")
print(f"Expected Votes={Exp_Votes}")
print("Null Hypothesis is that the sample data follows the given distribution.")
print("Alternate Hypothesis is that the sample data does not follows the given distribution.")
alpha=0.05
chi2_stats,p_value=scipy.stats.chisquare(Obs_Votes,Exp_Votes)
if p_value>alpha:
  print("Failed to reject null hypothesis for level of significance= "+str(alpha))
else:
  print("NUll hypothesis is rejected for level of significance= "+str(alpha))
```

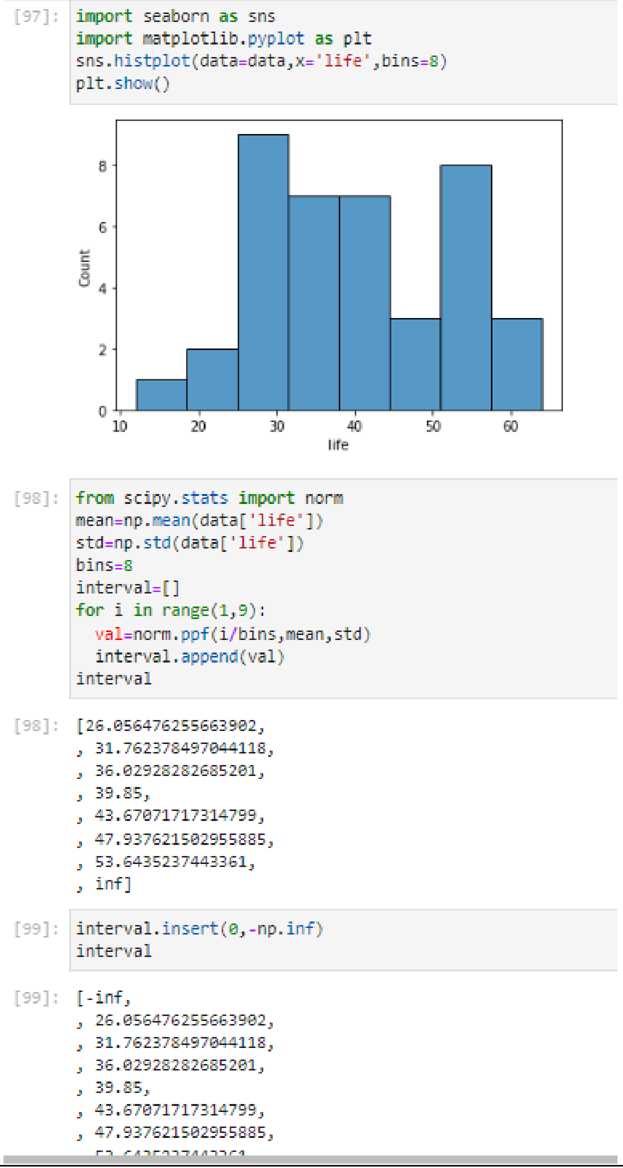
```
Observed Votes=[57 26 40]

Expected Votes=[51. 32. 40.]

Null Hypothesis is that the sample data follows the given distribution.

Alternate Hypothesis is that the sample data does not follows the given distribution.

Failed to reject null hypothesis for level of significance= 0.05
```

Question 7: A bulb manufacturer wants to know whether the life of the bulbs follows the normal distribution. Forty bulbs are randomly sampled, and their life, in months, are observed.



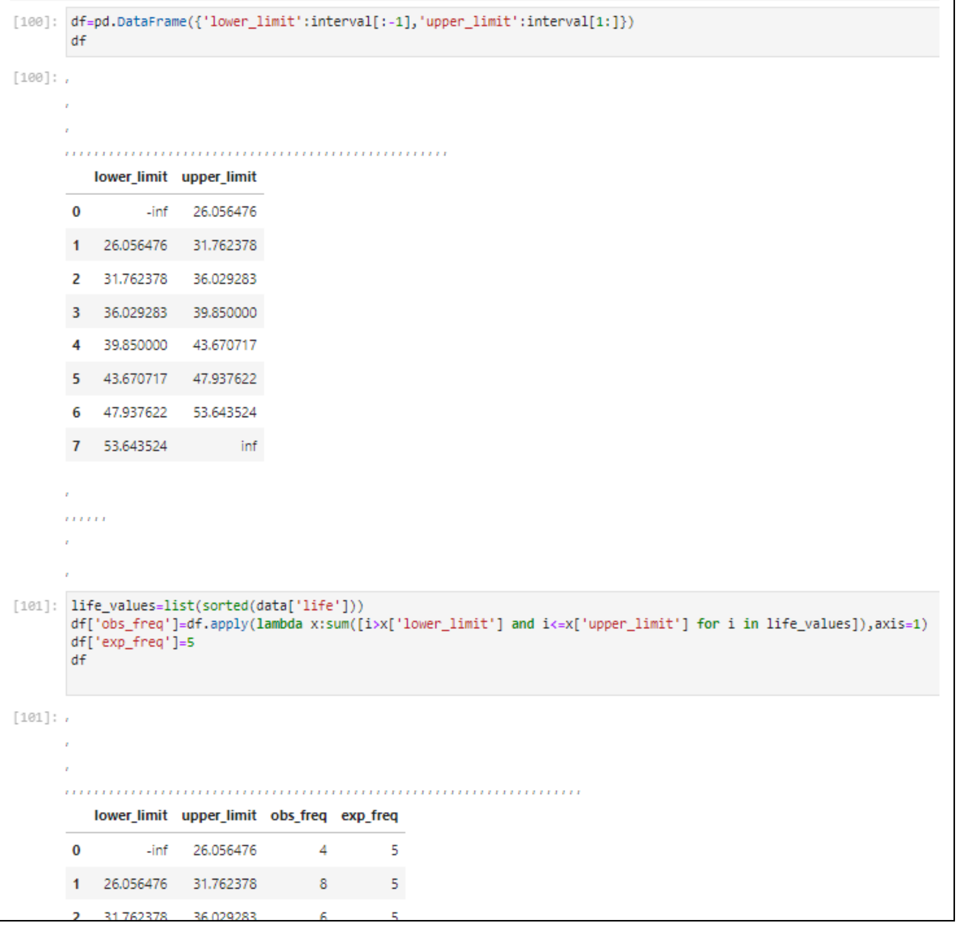
```
]: path = "/content/drive/MyDrive/SDS dataset/bulb_life.csv"
   data = pd.read_csv(path)
   data
```

```
]: ,
   ,
   ,
   ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,

       bulb  life
   0      1    31
   1      2    33
   2      3    34
   3      4    51
   4      5    24
   5      6    41
   6      7    58
   7      8    53
   8      9    27
   9     10    52
   10    11    40
   11    12    47
   12    13    37
   13    14    27
   14    15    31
   15    16    34
   16    17    34
   17    18    43
   18    19    55
```



```
[97]: import seaborn as sns
      import matplotlib.pyplot as plt
      sns.histplot(data=data,x='life',bins=8)
      plt.show()
```

```
[98]: from scipy.stats import norm
      mean=np.mean(data['life'])
      std=np.std(data['life'])
      bins=8
      interval=[]
      for i in range(1,9):
        val=norm.ppf(i/bins,mean,std)
        interval.append(val)
      interval

[98]: [26.056476255663902,
      , 31.762378497044118,
      , 36.02928282685201,
      , 39.85,
      , 43.67071717314799,
      , 47.937621502955885,
      , 53.6435237443361,
      , inf]

[99]: interval.insert(0,-np.inf)
      interval

[99]: [-inf,
      , 26.056476255663902,
      , 31.762378497044118,
      , 36.02928282685201,
      , 39.85,
      , 43.67071717314799,
      , 47.937621502955885,
      , 53.6435237443361,
```



[100]:
```
df=pd.DataFrame({'lower_limit':interval[:-1],'upper_limit':interval[1:]})
df
```

[100]: ,

,

,

,,

	lower_limit	upper_limit
0	-inf	26.056476
1	26.056476	31.762378
2	31.762378	36.029283
3	36.029283	39.850000
4	39.850000	43.670717
5	43.670717	47.937622
6	47.937622	53.643524
7	53.643524	inf

,

,,,,,,

,

,

[101]:
```
life_values=list(sorted(data['life']))
df['obs_freq']=df.apply(lambda x:sum([i>x['lower_limit'] and i<=x['upper_limit'] for i in life_values]),axis=1)
df['exp_freq']=5
df
```

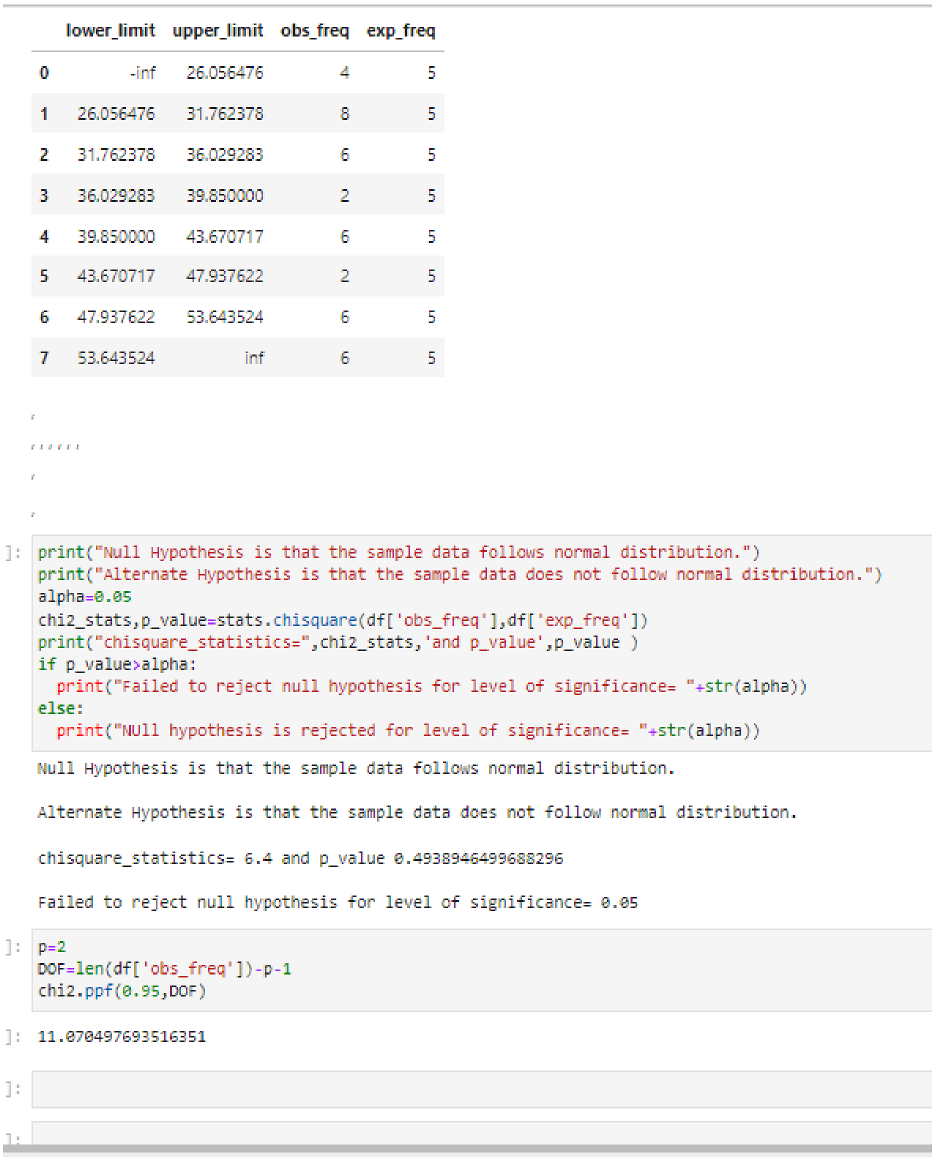
[101]: ,

,

,

,,

	lower_limit	upper_limit	obs_freq	exp_freq
0	-inf	26.056476	4	5
1	26.056476	31.762378	8	5
2	31.762378	36.029283	6	5



	lower_limit	upper_limit	obs_freq	exp_freq
0	-inf	26.056476	4	5
1	26.056476	31.762378	8	5
2	31.762378	36.029283	6	5
3	36.029283	39.850000	2	5
4	39.850000	43.670717	6	5
5	43.670717	47.937622	2	5
6	47.937622	53.643524	6	5
7	53.643524	inf	6	5

```
'
''''''
'
'
```

```
]: print("Null Hypothesis is that the sample data follows normal distribution.")
   print("Alternate Hypothesis is that the sample data does not follow normal distribution.")
   alpha=0.05
   chi2_stats,p_value=stats.chisquare(df['obs_freq'],df['exp_freq'])
   print("chisquare_statistics=",chi2_stats,'and p_value',p_value )
   if p_value>alpha:
     print("Failed to reject null hypothesis for level of significance= "+str(alpha))
   else:
     print("NUll hypothesis is rejected for level of significance= "+str(alpha))
```

```
Null Hypothesis is that the sample data follows normal distribution.

Alternate Hypothesis is that the sample data does not follow normal distribution.

chisquare_statistics= 6.4 and p_value 0.4938946499688296

Failed to reject null hypothesis for level of significance= 0.05
```

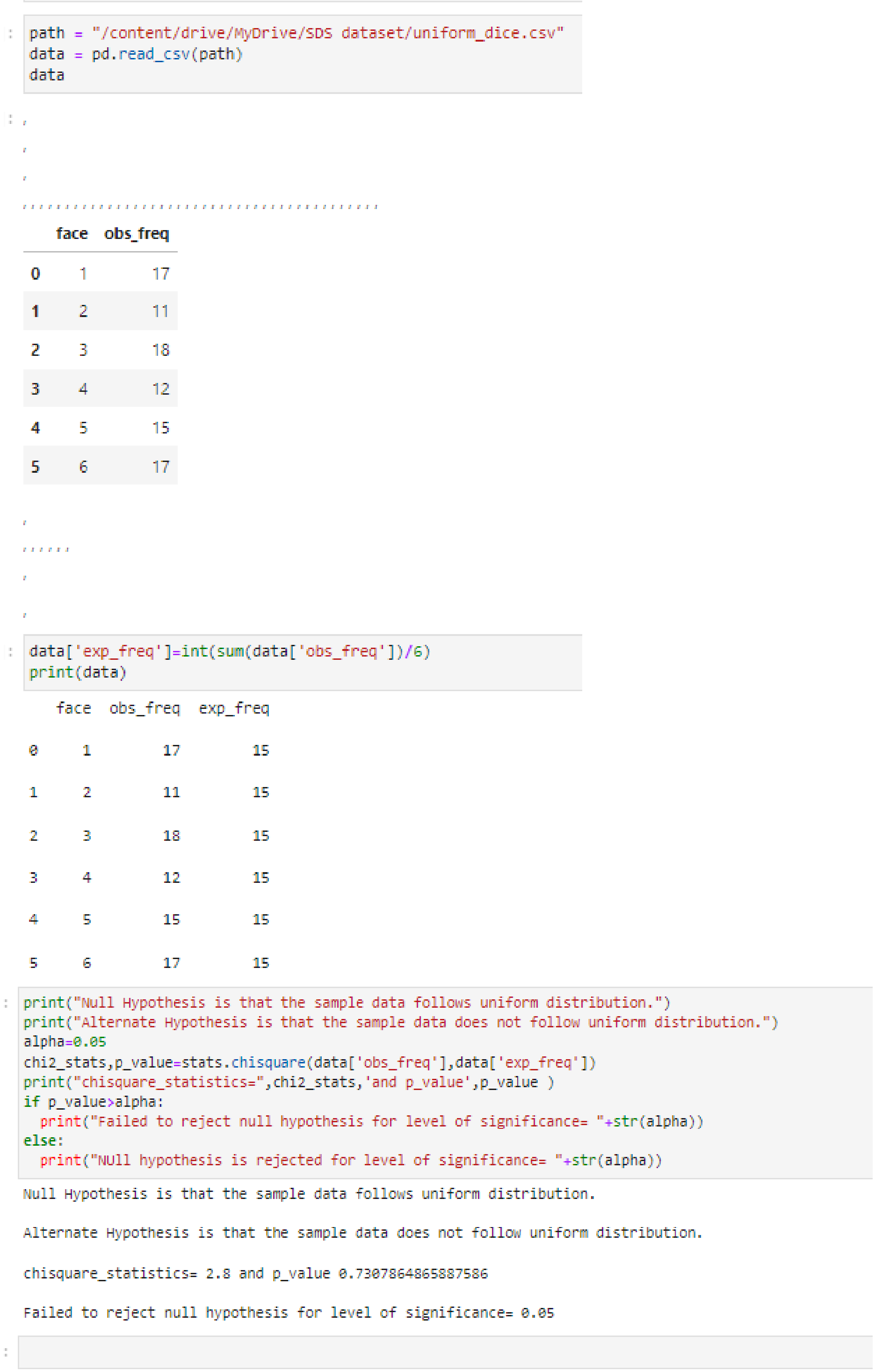
```
]: p=2
   DOF=len(df['obs_freq'])-p-1
   chi2.ppf(0.95,DOF)
```

```
]: 11.070497693516351
```

```
]:
```

```
]:
```

Question 8: Check whether the dice is unbiased. It is tossed 90 times and the counts of outcomes are given in table.



```
path = "/content/drive/MyDrive/SDS dataset/uniform_dice.csv"
data = pd.read_csv(path)
data
```

```
,
,
,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
```

	face	obs_freq
0	1	17
1	2	11
2	3	18
3	4	12
4	5	15
5	6	17

```
,
,,,,,,
,
,
```

```
data['exp_freq']=int(sum(data['obs_freq'])/6)
print(data)
```

```
   face  obs_freq  exp_freq
0     1        17        15
1     2        11        15
2     3        18        15
3     4        12        15
4     5        15        15
5     6        17        15
```

```
print("Null Hypothesis is that the sample data follows uniform distribution.")
print("Alternate Hypothesis is that the sample data does not follow uniform distribution.")
alpha=0.05
chi2_stats,p_value=stats.chisquare(data['obs_freq'],data['exp_freq'])
print("chisquare_statistics=",chi2_stats,'and p_value',p_value )
if p_value>alpha:
  print("Failed to reject null hypothesis for level of significance= "+str(alpha))
else:
  print("NUll hypothesis is rejected for level of significance= "+str(alpha))
```

```
Null Hypothesis is that the sample data follows uniform distribution.

Alternate Hypothesis is that the sample data does not follow uniform distribution.

chisquare_statistics= 2.8 and p_value 0.7307864865887586

Failed to reject null hypothesis for level of significance= 0.05
```

Conclusion	

Signature of Faculty