Assessment – 1

Name: Majjiga Jaswanth

RollNo: 20BCD7171

Question-1:

Assessment – I

1. a. Apply data transformation for the categorical attributes of Car Evaluation Data Set into numerical values.

Attribute	Transformation	
	From	To
buying	vhigh, high, med, low	3, 2, 1, 0
maint	vhigh, high, med, low	0, 1, 2, 3
lug_boot	small, med, big	0, 1, 2
safety	low, med, high	0, 1, 2

b. Draw a pie chart to represent the class distribution of car acceptance.

Code:

```
[14] import pandas as pd
     import numpy as np
     df=pd.read_csv("car.csv")
     df
     df.head()
```

	vhigh	vhigh.1	2	2.1	small	low	unacc
0	vhigh	vhigh	2	2	small	med	unacc
1	vhigh	vhigh	2	2	small	high	unacc
2	vhigh	vhigh	2	2	med	low	unacc
3	vhigh	vhigh	2	2	med	med	unacc
4	vhigh	vhigh	2	2	med	high	unacc

```
df.dtypes
```

```
vhigh      object
vhigh.1    object
2          object
2.1        object
small      object
low        object
unacc      object
dtype: object
```

```
obj_df = df.select_dtypes(include=['object']).copy()
obj_df.head()
```

	vhigh	vhigh.1	2	2.1	small	low	unacc
0	vhigh	vhigh	2	2	small	med	unacc
1	vhigh	vhigh	2	2	small	high	unacc
2	vhigh	vhigh	2	2	med	low	unacc
3	vhigh	vhigh	2	2	med	med	unacc
4	vhigh	vhigh	2	2	med	high	unacc

```
[17] obj_df[obj_df.isnull().any(axis=1)]
```

vhigh	vhigh.1	2	2.1	small	low	unacc

Question-2:

2. Develop user defined function to calculate chi-square correlation test for Nominal Data to decide the following two attributes **maint** and **class** are independent or dependent.

Note: Level of significance:0.05.

df	0.5	0.10	0.05	0.02	0.01	0.001
1	0.455	2.706	3.841	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.517

Code:

```
import pandas as pd
import numpy as np
def chi(observed,expected):
  chi=list()
  for i in observed:
     for k,j in observed.iterrows():
       if(i!='TOTAL' and k!='TOTAL'):
         print(i)
         print(k)
         print(observed[i][k])
         print(expected[i][k])
         a=observed[i][k]-expected[i][k]
         a=float(a*a/expected[i][k])
  if a<3.841:
    observed[i][k]='I'
```

```
  else:
    observed[i][k]='D'
    print(observed)

#data Taken from the assignment 6#
DATA={'HIGH SCHOOL':[90,60],'BACHELORS':[84,64],'MASTERS':[76,73],'PH.D
':[66,83]}
DATA=pd.DataFrame(DATA,index=pd.Index(['Female','Male']))
DATA['TOTAL']=DATA.sum(axis=1)
DATA.loc['TOTAL']=DATA.sum(axis=0)
DATA2=DATA.copy()
for i in DATA2:
  for k,j in DATA.iterrows():
    if(i!='TOTAL' and k!='TOTAL'):
       DATA2[i][k]=DATA[i]['TOTAL']*DATA.loc[k]['TOTAL']/DATA['TOTAL'][
'TOTAL']

print(DATA)
print(DATA2)
chi(DATA,DATA2)
```

```python
import pandas as pd
import numpy as np
def chi(observed,expected):
  chi=list()
  for i in observed:
     for k,j in observed.iterrows():
       if(i!='TOTAL' and k!='TOTAL'):
         print(i)
         print(k)
         print(observed[i][k])
         print(expected[i][k])
         a=observed[i][k]-expected[i][k]
         a=float(a*a/expected[i][k])
  if a<3.841:
    observed[i][k]='I'
  else:
    observed[i][k]='D'
    print(observed)
```

```python
#data Taken from the assignment 6#
DATA={'HIGH SCHOOL':[90,60],'BACHELORS':[84,64],'MASTERS':[76,73],'PH.D':[66,83]}
DATA=pd.DataFrame(DATA,index=pd.Index(['Female','Male']))
DATA['TOTAL']=DATA.sum(axis=1)
DATA.loc['TOTAL']=DATA.sum(axis=0)
DATA2=DATA.copy()
for i in DATA2:
  for k,j in DATA.iterrows():
    if(i!='TOTAL' and k!='TOTAL'):
       DATA2[i][k]=DATA[i]['TOTAL']*DATA.loc[k]['TOTAL']/DATA['TOTAL']['TOTAL']

print(DATA)
print(DATA2)
chi(DATA,DATA2)
```

```
        HIGH SCHOOL  BACHELORS  MASTERS  PH.D  TOTAL
Female           90         84       76    66    316
Male             60         64       73    83    280
TOTAL           150        148      149   149    596
        HIGH SCHOOL  BACHELORS  MASTERS  PH.D  TOTAL
Female           79         78       79    79    316
Male             70         69       70    70    280
TOTAL           150        148      149   149    596
HIGH SCHOOL
Female
90
79
HIGH SCHOOL
Male
60
70
BACHELORS
Female
84
78
BACHELORS
Male
64
69
MASTERS
Female
76
79
MASTERS
Male
73
70
```

Question-3:

3. Develop user defined function to calculate dissimilarity matrix for the 10 random objects of Car Evaluation Data set by considering following categorical attributes: buying, maint, doors, persons, lug_boot, safety.

Code:

```
import pandas as pd
import numpy as np

df = pd.read_csv("car.csv")


df.columns = ['buying','maint','doors','persons','lug_boot','safety','a
cc']
```

```
b=df [ 'buying' ]
m=df [ 'maint' ]
d=df [ 'doors' ]
p=df [ 'persons' ]
l=df [ 'lug_boot' ]
s=df [ 'safety' ]

import random
v1=random. choice (b)
v2=random. choice (m)
v3=random. choice (d)
v4=random. choice (p)
v5=random. choice (l)
v6=random. choice (s)
v7=random. choice (b)
v8=random. choice (m)
v9=random. choice (d)
v10=random. choice (p)

arr=np.array([[v1,v2,v3],
 [v4,v5,v6],
 [v7,v8,v1]])

l1=[]
l2=[]
for i in range(0,len(arr)):
 for j in range(0,len(arr)):
        Aij = i
        Bij = j
        l1.append(Aij)
        l2.append(Bij)
print("Dissimarity matrix \n ",np.array([l1,l2]))
```

output:

```
Dissimarity matrix
  [[0 0 0 1 1 1 2 2 2]
 [0 1 2 0 1 2 0 1 2]]
```