Darshan UNIVERSITY
योग: कर्मसु कौशलम्

(https://www.darshan.ac.in/)

Data Mining

Lab - 4

Name : Krish Gohel

Enr No : 22010101060

Roll No. : 216

batch : A6

Part -1

1) Write a python program to compute distance between Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

(a) Compute the Euclidean distance between the two objects.
(b) Compute the Manhattan distance between the two objects.
(c) Compute the Minkowski distance between the two objects, using q = 3.
(d) Compute the supremum distance between the two objects.

In [6]:
```
ob1 = (22,1,42,10)
ob2 = (20,0,36,8)

ans = 0.0
for o1,o2 in zip(ob1,ob2):
    ans = ans + (o1-o2)**2
print('Euclidean distance : ',ans*0.5)

print('-------------------------')

ans = 0.0
for o1,o2 in zip(ob1,ob2):
    ans += abs(o1-o2)
print('Manhattan distance : ',ans)

print('-------------------------')

ans = 0.0
q = 3
for o1,o2 in zip(ob1,ob2):
    ans += abs(o1-o2)**q
print('Minkowski distance :',ans**(1/q))

print('-------------------------')

for_max=[]
for o1,o2 in zip(ob1,ob2):
    for_max.append(o1-o2)
print('Supernum : ',max(for_max))
```

```
Euclidean distance :  22.5
-------------------------
Manhattan distance :  11.0
-------------------------
Minkowski distance : 6.153449493663682
-------------------------
Supernum :  6
```

In [11]:
```
ob1 = (22, 1, 42, 10)
ob2 = ob2 = (20,0,36,8)
sum = 0

for i in range(0,len(ob1)):
    sum += (ob1[i] - ob2[i])**2
print('Euclidian result : ',sum**.5)

print("----------------------")
sum = 0
for i in range(0,len(ob1)):
    sum += abs((ob1[i] - ob2[i]))
print('manhattan result : ',sum)

print("----------------------")
sum = 0
q = 3
sum = 0
for i in range(0,len(ob1)):
    sum += (abs(ob1[i] - ob2[i]))**q
print('Minkowski result : ',sum**(1/q))

print("----------------------")

for_max=[]
for i in range(0,len(ob1)):
    x = abs(ob1[i]-ob2[i])
    for_max.append(x)
print('Supernum : ',max(for_max))
```

```
Euclidian result :  6.708203932499369
----------------------
manhattan result :  11
----------------------
Minkowski result :  6.153449493663682
----------------------
Supernum :  6
```

2) Perform Preprocessing on Titanic Data set Using Orange Tools

3) Kindly Perform Data Exploration on New Restaurant Data Set

Link -
https://github.com/guipsamora/pandas_exercises/blob/master/01_Getting_%26_Knowing_Your_D
(https://github.com/guipsamora/pandas_exercises/blob/master/01_Getting_%26_Knowing_Your_

In [ ]:

PART - 2

In [2]:
```
import pandas as pd
```

In [33]:
```
data = pd.read_csv("titanic.csv")
```

1) First, you need to read the titanic dataset from local disk and display Last five records

In [4]:
```
data.tail(5)
```

Out[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	NaN
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN

2) Handle Missing Values in data set [use dropna(), fillna(), and interpolate]

In [7]:
```
data.isnull().sum()
```

Out[7]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

In [10]:
```
data.fillna({'Age':0,'Cabin':False,'Embarked':'S'})
```

Out[10]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500
...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	0.0	1	2	W./C. 6607	23.4500
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500

891 rows × 12 columns

In [18]:
```
data.interpolate(axis=0,method='polynomial',order=3)
```

Out[18]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	I
0	1	0	3	Braund, Mr. Owen Harris	male	22.000000	1	0	A/5 21171	7.2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.000000	1	0	PC 17599	71.2
2	3	1	3	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282	7.9
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.000000	1	0	113803	53.1
4	5	0	3	Allen, Mr. William Henry	male	35.000000	0	0	373450	8.0
...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536	13.0
887	888	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053	30.0
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	20.315232	1	2	W./C. 6607	23.4
889	890	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369	30.0
890	891	0	3	Dooley, Mr. Patrick	male	32.000000	0	0	370376	7.7

891 rows × 12 columns

3) Write programs to perform the following tasks of preprocessing.

Equal Width Binning

Equal Frequency/Depth Binning

In [27]:
```
import pandas as pd
import numpy as np

data = [5,10,11,13,15,35,50,55,72,92,204,215]

df = pd.DataFrame(data,columns=['value'])

num_bins = 3

bin_edges = np.linspace(df['value'].min(),df['value'].max(),num_bins+1)
print(bin_edges)
df['Equal_Width_Bin'] = pd.cut(df['value'],bins=bin_edges,labels=['b1','b2','b

print("Equal Width Binning \n",df)
```

```
[  5.  75. 145. 215.]
Equal Width Binning
     value Equal_Width_Bin
0       5              b1
1      10              b1
2      11              b1
3      13              b1
4      15              b1
5      35              b1
6      50              b1
7      55              b1
8      72              b1
9      92              b2
10    204              b3
11    215              b3
```

In [4]:
```
import pandas as pd

def EFB(data,nbin):
    df = pd.DataFrame(data,columns=["value"])
    df['bin']= pd.qcut(df["value"],q=nbin,labels=['b1','b2','b3'])
    binn_data = df.groupby('bin')['value']
    return binn_data

print('equal frequncy binning : \n')

data = [5,10,11,13,15,35,50,55,72,92,204,215]

nbin = 3

bd = EFB(data,nbin)

for bin_id,val in bd:
    print(bin_id,val)
```

```
equal frequncy binning :

b1 0     5
1    10
2    11
3    13
Name: value, dtype: int64
b2 4    15
5    35
6    50
7    55
Name: value, dtype: int64
b3 8      72
9      92
10    204
11    215
Name: value, dtype: int64
```

4) Apply Scaling to AGE attribute with min max, decimal scaling and z score.

In [20]:
```
import pandas as pd

def FL(n):
    count = 0
    while n != 0:
        count += 1
        n = int(n/10)
    return count

x = data['Age'].max()
y = FL(int(x))
V = x/10**y
V

data['Age'][0] = data['Age'][0]/10**y
data
```

```
C:\Users\Krish\AppData\Local\Temp\ipykernel_17820\2095767983.py:15: SettingWi
thCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/s
table/user_guide/indexing.html#returning-a-view-versus-a-copy (https://panda
s.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ver
sus-a-copy)
  data['Age'][0] = data['Age'][0]/10**y
```

Out[20]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	1	0	3	Braund, Mr. Owen Harris	male	0.22	1	0	A/5 21171	7.2500
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	0.38	1	0	PC 17599	71.2833
2	3	1	3	Heikkinen, Miss. Laina	female	0.26	0	0	STON/O2. 3101282	7.9250
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	0.35	1	0	113803	53.1000
4	5	0	3	Allen, Mr. William Henry	male	0.35	0	0	373450	8.0500
...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	0.27	0	0	211536	13.0000
887	888	1	1	Graham, Miss. Margaret Edith	female	0.19	0	0	112053	30.0000
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500
889	890	1	1	Behr, Mr. Karl Howell	male	0.26	0	0	111369	30.0000
890	891	0	3	Dooley, Mr. Patrick	male	0.32	0	0	370376	7.7500

891 rows × 12 columns

In [41]:
```
import pandas as pd

def FL(n):
    count = 0
    while n != 0:
        count += 1
        n = int(n/10)
    return count

x = data['Age'].min()
y = FL(int(x))
V = x/10**y
V

data['Age'][0] = data['Age'][0]/10**y
data
```

```
C:\Users\Krish\AppData\Local\Temp\ipykernel_17820\2892590180.py:15: SettingWi
thCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/s
table/user_guide/indexing.html#returning-a-view-versus-a-copy (https://panda
s.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ver
sus-a-copy)
  data['Age'][0] = data['Age'][0]/10**y
```

Out[41]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500
...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500

891 rows × 12 columns

In [29]:
```
import pandas as pd
import numpy as np

m = data['Age'].mean()
sd = np.std(data['Age'])

v = 215
V = (v-m)/sd
V

for i in range(0,len(data)):
    data['Age'][i] = (data['Age'][i] - m)/sd

data
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  data['Age'][i] = (data['Age'][i] - m)/sd
C:\Users\Krish\AppData\Local\Temp\ipykernel_17820\810244196.py:12: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  data['Age'][i] = (data['Age'][i] - m)/sd
C:\Users\Krish\AppData\Local\Temp\ipykernel_17820\810244196.py:12: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame
```