Darshan UNIVERSITY
योग: कर्मसु कौशलम्

Data Mining

Lab - 4

Name : Vyas Bhagyesh Y.

Div : 5-B

Batch : 2

Roll-No : 423

Enrollment No : 23010101662

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 [2]:

```
import math as m
x=(22,1,42,10)
y=(20,0,36,8)
sum=0
for i in range(len(x)):
    sum+=(x[i]-y[i])**2

print(f"Euclidiean distance :{m.sqrt(sum)}")
```

```
Euclidiean distance :6.708203932499369
```

In [4]:

```
import math as m
x=(22,1,42,10)
y=(20,0,36,8)
sum=0
for i in range(len(x)):
```

```
    sum+=abs(x[i]-y[i])

print(f"Manhattan distance :{sum}")
```

```
Manhattan distance :11
```

In [20]:

```
import math as m
x=(22,1,42,10)
y=(20,0,36,8)
sum=0
q=3
for i in range(len(x)):
    sum+=abs(x[i]-y[i])**q

print(f"Minkowski distance :{sum**(1/q)}")
```

```
Minkowski distance :6.153449493663682
```

In [21]:

```
import math as m
x=(22,1,42,10)
y=(20,0,36,8)
sum=0
maxList=[abs(x[i]-y[i]) for i in range(len(x))]
# for i in range(len(x)):
#     maxList.append(abs(x[i]-y[i]))

print(f"Supremum distance :{max(maxList)}")
# print(maxList)
```

```
Supremum distance :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_Data/Chipotle/Ex

In [ ]:

PART - 2

In [2]:

```
import pandas as pd
import numpy as np
```

In [20]:

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

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

In [21]:

```
df.tail()
```

Out[21]:

Out[21]:

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

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

In [5]:

```
df.count()
```

Out[5]:

```
PassengerId    891
Survived       891
Pclass         891
Name           891
Sex            891
Age            714
SibSp          891
Parch          891
Ticket         891
Fare           891
Cabin          204
Embarked       889
dtype: int64
```

In [6]:

```
df.isnull()
```

Out[6]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	False	False	False	False	False	False	False	False	False	False	True	False
1	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	True	False
3	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	True	False
...	...	...	...	...	...	...	...	...	...	...	...	...
886	False	False	False	False	False	False	False	False	False	False	True	False
887	False	False	False	False	False	False	False	False	False	False	False	False
888	False	False	False	False	False	True	False	False	False	False	True	False
889	False	False	False	False	False	False	False	False	False	False	False	False
890	False	False	False	False	False	False	False	False	False	False	True	False

891 rows × 12 columns

In [6]:

```
print(df.isnull().sum())
```

```
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 [8]:

```
print(df.Age.isnull().value_counts())
```

```
Age
False    714
True     177
Name: count, dtype: int64
```

In [7]:

```
df2=df.dropna()
df2
```

Out[7]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S
...	...	...	...	...	...	...	...	...	...	...	...	...
871	872	1	1	Beckwith, Mrs. Richard Leonard (Sallie Monypeny)	female	47.0	1	1	11751	52.5542	D35	S
872	873	0	1	Carlsson, Mr. Frans Olof	male	33.0	0	0	695	5.0000	B51 B53 B55	S
879	880	1	1	Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)	female	56.0	0	1	11767	83.1583	C50	C
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C

183 rows × 12 columns

In [8]:

```
df3=df.fillna({'Age':0})
df3
```

Out[8]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	0.0	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

In [10]:

```
df3=df3.fillna({'Cabin':'Not Allocated'})
df3
```

Out[10]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Not Allocated	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Not Allocated	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Not Allocated	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	Not Allocated	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	0.0	1	2	W./C. 6607	23.4500	Not Allocated	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	Not Allocated	Q

891 rows × 12 columns

In [11]:

```
df3=df3.fillna({'Age':0,'Cabin':'Not Allocated','Embarked':'DEFAULT'})
df3
```

Out[11]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Not Allocated	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Not Allocated	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Not Allocated	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	Not Allocated	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	0.0	1	2	W./C. 6607	23.4500	Not Allocated	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	Not Allocated	Q

891 rows × 12 columns

In [9]:

```
df3.count()
```

Out[9]:

```
PassengerId    891
Survived       891
Pclass         891
Name           891
Sex            891
Age            891
SibSp          891
Parch          891
Ticket         891
Fare           891
Cabin          204
Embarked       889
dtype: int64
```

In [10]:

```
print(df3.isnull().sum())
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
```

```
Age              0
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

In [21]:

```
print(df3.Embarked.isnull().value_counts())
```

```
Embarked
False    891
Name: count, dtype: int64
```

In [22]:

```
df.count()
```

Out[22]:

```
PassengerId    891
Survived       891
Pclass         891
Name           891
Sex            891
Age            714
SibSp          891
Parch          891
Ticket         891
Fare           891
Cabin          204
Embarked       889
dtype: int64
```

In [15]:

```
df4=df.fillna({'Age':df.Age.mean()})
df4
```

Out[15]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.000000	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.000000	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.000000	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.000000	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.699118	1	2	W./C. 6607	23.4500	NaN	S

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
889	890	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.000000	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

In [16]:

```
df5=df.fillna({'Age':df.Age.mode()[0]})
df5
```

Out[16]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	24.0	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

In [17]:

```
df6=df.Age.interpolate(mathod='limit',linear_direction='Forward',axis=0)
df6
```

Out[17]:

```
0      22.0
1      38.0
2      26.0
3      35.0
4      35.0
       ...
886    27.0
887    19.0
888    22.5
889    26.0
890    32.0
Name: Age, Length: 891, dtype: float64
```

In [26]:

```
df7=df.fillna({'Age':df.Age.interpolate(mathod='limit',linear_direction='Forward',axis=0
```

```
)})
df7
```

Out[26]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	22.5	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

In [ ]:

```

```

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

Equal Width Binning
Equal Frequency/Depth Binning

In [52]:

```
data = [5,10,11,13,15,35,50,55,72,92,204,215]
df = pd.DataFrame(data, columns=['Values'])
num_bins = 3
bin_edges = np.linspace(df['Values'].min(),df['Values'].max(),num_bins+1)
df['Equal_Width_Bin'] = pd.cut(df['Values'],bins=bin_edges,labels=[1,2,3], include_lowes
t=True)
print("Equal Width Binning:\n" ,df)
```

```
Equal Width Binning:
     Values Equal_Widht_Bin
0         5               1
1        10               1
2        11               1
3        13               1
4        15               1
5        35               1
6        50               1
7        55               1
8        72               1
9        92               2
10      204               3
```

```
11     215                3
```

In [54]:

```
#Equal frequency(depth) binning

no_of_data = len(data)
points_in_bin = no_of_data/num_bins

for i in range(0,len(data),int(points_in_bin)):
    print(data[i:i+int(points_in_bin)])
```

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

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

In [31]:

```
df8=df7
new_min=1
new_max=10
df8['Age Normalized']=(((df7.Age-df7.Age.min())/(df7.Age.max()-df7.Age.min()))*(new_max-new_min))+new_min
df8['Age Normalized']
```

Out[31]:

```
0      3.440563
1      5.250063
2      3.892938
3      4.910782
4      4.910782
         ...
886    4.006032
887    3.101282
888    3.497110
889    3.892938
890    4.571500
Name: Age Normalized, Length: 891, dtype: float64
```

In [32]:

```
df8['Age Decimal Scaling']=df7['Age']/10**len(str(int(df7['Age'].max())))
# print(len(str(int(df7.Age.max()))))
df8['Age Decimal Scaling']
```

Out[32]:

```
0      0.220
1      0.380
2      0.260
3      0.350
4      0.350
       ...
886    0.270
887    0.190
888    0.225
889    0.260
890    0.320
Name: Age Decimal Scaling, Length: 891, dtype: float64
```

In [33]:

```
df8['Age Z-Score']=(df7['Age']-df7['Age'].mean())/df7['Age'].std()
df8
```

Out[33]:

	PassengerId PassengerId	Survived Survived	Pclass Pclass	Name Name	Sex Sex	Age Age	SibSp SibSp	Parch Parch	Ticket Ticket	Fare Fare	Cabin Cabin	Embarked Embarked	Age Normalized Age Normalized
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	3.44056
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	5.25006
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	3.89293
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	4.91078
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	4.91078
...	...	...	...	...	...	...	...	...	...	...	...	...	.
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S	4.00603
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S	3.10128
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	22.5	1	2	W./C. 6607	23.4500	NaN	S	3.49711
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C	3.89293
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q	4.57150

891 rows × 15 columns

In [34]:

```
df8[['Age','Age Normalized','Age Decimal Scaling','Age Z-Score']].corr()
```

Out[34]:

	Age	Age Normalized	Age Decimal Scaling	Age Z-Score
Age	1.0	1.0	1.0	1.0
Age Normalized	1.0	1.0	1.0	1.0
Age Decimal Scaling	1.0	1.0	1.0	1.0
Age Z-Score	1.0	1.0	1.0	1.0