Darshan UNIVERSITY

योग: कर्मसु कौशलम्

Data Mining

Lab - 1

Introduction to Pandas Library Function:

Step-1 Import the pandas Libraries

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

Step-2 Import the dataset from this:....

In [41]:
```
file = "titanic.csv"
```

Step-3 Read csv or excel File

In [42]:
```
df = pd.read_csv(file)
```

Step-4 Print Data from csv or excel File

In [7]:
```
df
```

Out[7]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.25
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.28
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.92
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.10
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.05
...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75

891 rows × 12 columns

Step-5 See the First 10 Rows

In [8]:
```
df.head(10)
```

Out[8]:

	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
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708

Step-6 See the Last 10 Rows

In [9]:
```
df.tail(10)
```

Out[9]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
881	882	0	3	Markun, Mr. Johann	male	33.0	0	0	349257	7.
882	883	0	3	Dahlberg, Miss. Gerda Ulrika	female	22.0	0	0	7552	10.
883	884	0	2	Banfield, Mr. Frederick James	male	28.0	0	0	C.A./SOTON 34068	10.
884	885	0	3	Sutehall, Mr. Henry Jr	male	25.0	0	0	SOTON/OQ 392076	7.
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.

Step-7 Data type of each columns

In [10]:
```
df.dtypes
```

Out[10]:
```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

Step-8 Display Summary Information

In [11]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

Step-9 Access a specific column

In [12]:
```
df["Ticket"]
```

Out[12]:
```
0             A/5 21171
1              PC 17599
2      STON/O2. 3101282
3                113803
4                373450
              ...
886              211536
887              112053
888          W./C. 6607
889              111369
890              370376
Name: Ticket, Length: 891, dtype: object
```

Step-10 Access rows by their integer location

In [18]:
```
df.iloc[100]
```

Out[18]:
```
PassengerId                                  102
Survived                                       0
Pclass                                         3
Name           Petroff, Mr. Pastcho ("Pentcho")
Sex                                         male
Age                                          NaN
SibSp                                          0
Parch                                          0
Ticket                                    349215
Cabin                                        NaN
Embarked                                       S
Name: 101, dtype: object
```

Step-11 Delete a specific Column

In [20]:
```
df = df.drop(columns= ["Fare"])
df
```

Out[20]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Cabi
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	Na
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	Na
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	C12
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	Na
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	Na
...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	Na
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	B4
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	Na
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	C14
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	Na

890 rows × 11 columns

Step-12 Create a new Column

In [22]:
```
df["Fare"] = "1"
```

In [28]:
```
df
```

Out[28]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Cabi
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	Na
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	Na
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	C12
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	Na
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	Na
...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	Na
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	B4
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	Na
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	C14
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	Na

890 rows × 12 columns

Step-13 Perform Condition Selection on DataFrame

In [31]:
```
data = df[df['Age'] > 65]
print(data)
```

```
     PassengerId  Survived  Pclass                                     Name  \
33            34         0       2                    Wheadon, Mr. Edward H
96            97         0       1                  Goldschmidt, Mr. George B
116          117         0       3                       Connors, Mr. Patrick
493          494         0       1                    Artagaveytia, Mr. Ramon
630          631         1       1  Barkworth, Mr. Algernon Henry Wilson
672          673         0       2               Mitchell, Mr. Henry Michael
745          746         0       1              Crosby, Capt. Edward Gifford
851          852         0       3                       Svensson, Mr. Johan

      Sex   Age  SibSp  Parch      Ticket Cabin Embarked Fare
33   male  66.0      0      0  C.A. 24579   NaN        S    1
96   male  71.0      0      0    PC 17754    A5        C    1
116  male  70.5      0      0      370369   NaN        Q    1
493  male  71.0      0      0    PC 17609   NaN        C    1
630  male  80.0      0      0       27042   A23        S    1
672  male  70.0      0      0  C.A. 24580   NaN        S    1
745  male  70.0      1      1   WE/P 5735   B22        S    1
851  male  74.0      0      0      347060   NaN        S    1
```

Step-14 Compute the sum of value

In [32]:
```
data = df['Survived'].sum()
print(data)
```

```
341
```

Step-15 Compute the mean of value

In [33]:
```
data = df['Age'].mean()
print(data)
```

```
29.687475455820476
```

Step-16 Count non-null value (column)

In [38]:
```
data =  df['Age'].count()
print(data)
```

```
713
```

Step-17 Find Minimun or Maximum values

In [50]:
```
minAge = df['Age'].min()
maxAge = df['Age'].max()
```

```
print(minAge)
print(maxAge)
```

```
0.42
80.0
```