



Data Mining

Lab - 1

140 | Vishal Baraiya |
23010101014

Introduction to Pandas Library Function:

Step-1 Import the pandas Libraries

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

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

```
In [7]:
```

Step-3 Read csv or excel File

```
In [3]: df= pd.read_csv("titanic.csv")  
df
```

Out[3]:	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	7.2!
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0	1	0	PC 17599	71.2!
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9!
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1!
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0!
...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0!
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0!
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4!
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0!
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7!

891 rows × 12 columns



Step-4 Print Data from csv or excel File

```
In [5]: df.shape
```

```
Out[5]: (891, 12)
```

Step-5 See the First 10 Rows

```
In [6]: df.head(10)
```

Out[6]:	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	51.8621
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 [7]: `df.tail(10)`

Out[7]:

	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	Suttehall, 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 [9]: `df.dtypes`

```
Out[9]: 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 [12]: `df.describe()`

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Step-9 Access a specific column

In [21]: `# specifically colmun access karava mate`
`df["Name"]`

```
Out[21]: 0 Braund, Mr. Owen Harris
1 Cummings, Mrs. John Bradley (Florence Briggs Th...
2 Heikkinen, Miss. Laina
3 Futrelle, Mrs. Jacques Heath (Lily May Peel)
4 Allen, Mr. William Henry
...
886 Montvila, Rev. Juozas
887 Graham, Miss. Margaret Edith
888 Johnston, Miss. Catherine Helen "Carrie"
889 Behr, Mr. Karl Howell
890 Dooley, Mr. Patrick
Name: Name, Length: 891, dtype: object
```

In [8]: `df[["Name", "Age", "Parch", "Fare"]]`

	Name	Age	Parch	Fare
0	Braund, Mr. Owen Harris	22.0	0	7.2500
1	Cummings, Mrs. John Bradley (Florence Briggs Th...	38.0	0	71.2833
2	Heikkinen, Miss. Laina	26.0	0	7.9250
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	35.0	0	53.1000
4	Allen, Mr. William Henry	35.0	0	8.0500
...
886	Montvila, Rev. Juozas	27.0	0	13.0000
887	Graham, Miss. Margaret Edith	19.0	0	30.0000
888	Johnston, Miss. Catherine Helen "Carrie"	NaN	2	23.4500
889	Behr, Mr. Karl Howell	26.0	0	30.0000
890	Dooley, Mr. Patrick	32.0	0	7.7500

891 rows × 4 columns

Step-10 Access rows by their integer location

In [20]: `# specifically raw access karava mate
df.iloc[0]`

```
Out[20]: PassengerId      1
          Survived       0
          Pclass        3
          Name    Braund, Mr. Owen Harris
          Sex            male
          Age         22.0
          SibSp         1
          Parch        0
          Ticket     A/5 21171
          Fare        7.25
          Cabin        NaN
          Embarked      S
          Name: 0, dtype: object
```

In [9]: `df.iloc[[0, 2]]`

	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	7.250
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925



In [10]: `df.iloc[0:3] # Rows 0 to 2 (not including 3)`

	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	7.250
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



In [11]: `df[df['Age'] > 25]`

Out[11]:		PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
					Cumings, Mrs. John Bradley (Florence Briggs Th...						
1	2	1	1	1	Heikkinen, Miss. Laina	female	38.0	1	0	PC 17599	71.
2	3	1	3		Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	26.0	0	0	STON/O2. 3101282	7.
3	4	1	1		Allen, Mr. William Henry	male	35.0	1	0	113803	53.
4	5	0	3		McCarthy, Mr. Timothy J	male	35.0	0	0	373450	8.
6	7	0	1		Banfield, Mr. Frederick James	male	54.0	0	0	17463	51.
...
883	884	0	2		Rice, Mrs. William (Margaret Norton)	male	28.0	0	0	C.A./SOTON 34068	10.
885	886	0	3		Montvila, Rev. Juozas	female	39.0	0	5	382652	29.
886	887	0	2		Behr, Mr. Karl Howell	male	27.0	0	0	211536	13.
889	890	1	1		Dooley, Mr. Patrick	male	26.0	0	0	111369	30.
890	891	0	3							370376	7.

413 rows × 12 columns



In [18]: df.loc[883]

```
Out[18]: PassengerId      884
          Survived        0
          Pclass           2
          Name    Banfield, Mr. Frederick James
          Sex            male
          Age            28.0
          SibSp          0
          Parch          0
          Ticket       C.A./SOTON 34068
          Fare           10.5
          Cabin          NaN
          Embarked        S
          Name: 883, dtype: object
```

Step-11 Delete a specific Column

```
In [26]: # specific column delete karava mate
          df.drop("Embarked",axis="columns",inplace=True)
```

Step-12 Create a new Column

```
In [30]: df
```

Out[30]:	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	7.2!
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0	1	0	PC 17599	71.2!
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9!
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1!
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0!
...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0!
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0!
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4!
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0!
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7!

891 rows × 12 columns



```
In [31]: df["isCabin"] = ~ df["Cabin"].isnull()
```

```
In [32]: df
```

Out[32]:	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	7.2!
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0	1	0	PC 17599	71.2!
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9!
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1!
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0!
...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0!
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0!
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4!
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0!
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7!

891 rows × 12 columns



Step-13 Perform Condition Selection on DataFrame

```
In [33]: df[df[ "Age" ] > 30]
```

Out[33]:	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
				Cumings, Mrs. John Bradley (Florence Briggs Th...						
1	2	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	38.0	1	0	PC 17599	71.2833
3	4	1	1	Allen, Mr. William Henry	male	35.0	1	0	113803	53.1000
4	5	0	3	McCarthy, Mr. Timothy J	male	54.0	0	0	373450	8.0500
6	7	0	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	17463	51.8625
11	12	1	1	Vander Cruyssen, Mr. Victor	male	47.0	0	0	113783	26.5500
...	Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)	female	56.0	0	1	345765	9.0000
873	874	0	3	Markun, Mr. Johann	male	33.0	0	0	11767	83.1583
881	882	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	349257	7.8958
885	886	0	3	Dooley, Mr. Patrick	male	32.0	0	0	382652	29.1250
890	891	0	3						370376	7.7500

305 rows × 12 columns



In [34]: `df[df['Pclass'] == 3]`

Out[34]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171 7.
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282 7.
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450 8.
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877 8.
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909 21.
...
882	883	0	3	Dahlberg, Miss. Gerda Ulrika	female	22.0	0	0	7552 10.
884	885	0	3	Suttehall, 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.
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607 23.
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376 7.

491 rows × 12 columns



Step-14 Compute the sum of value

```
In [35]: df["Fare"].sum()
```

```
Out[35]: 28693.9493
```

Step-15 Compute the mean of value

```
In [36]: df["Fare"].mean()
```

```
Out[36]: 32.204207968574636
```

Step-16 Count non-null value (column)

```
In [37]: (~df.isnull()).sum()
```

```
Out[37]: PassengerId      891
Survived          891
Pclass            891
Name              891
Sex               891
Age              714
SibSp            891
Parch            891
Ticket           891
Fare             891
Cabin           204
isCabin         891
dtype: int64
```

Step-17 Find Minimum or Maximum values

```
In [38]: df["Fare"].min()
```

```
Out[38]: 0.0
```

```
In [39]: df["Fare"].max()
```

```
Out[39]: 512.3292
```

```
In [40]: df["Fare"].median()
```

```
Out[40]: 14.4542
```

```
In [41]: df["Fare"].mode()
```

```
Out[41]: 0    8.05
Name: Fare, dtype: float64
```

```
In [42]: df["Fare"].std()
```

Out[42]: 49.693428597180905

```
In [ ]: # generate a excel file  
df.to_excel("output.xlsx")
```