



Experiment No. 1
Exploratory Data Analysis

Student Name: Lipakshi
Branch: BE-CSE
Semester: 5th

UID: 20BCS5082
Section/Group: 607/B
Subject: Machine Learning Lab

1. **Aim:** We have to implement data analysis on the csv file named titanic train 1.csv. We will use pandas library as it is used for data cleaning and analysis.
2. **Software/Hardware Requirements:** Windows 7 & above version
3. **Tools to be used:**
 - Anaconda Navigator
 - Jupiter Notebook
4. **Implementation:**

```
In [3]: #reading the csv file in dataframe
df=pd.read_csv("F:\\SEM-5\\ML\\titanic-train.csv")
print(df)
```

	PassengerId	Survived	Pclass	\
0	1	0	3	
1	2	1	1	
2	3	1	3	
3	4	1	1	
4	5	0	3	
..	
886	887	0	2	
887	888	1	1	
888	889	0	3	
889	890	1	1	
890	891	0	3	

	Name	gender	Age	SibSp	\
0	Braund, Mr. Owen Harris	male	22.0	1	
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	
2	Heikkinen, Miss. Laina	female	26.0	0	
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	
4	Allen, Mr. William Henry	male	35.0	0	
..	
886	Montvila, Rev. Juozas	male	27.0	0	
887	Graham, Miss. Margaret Edith	female	19.0	0	
888	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	
889	Behr, Mr. Karl Howell	male	26.0	0	
890	Dooley, Mr. Patrick	male	32.0	0	

	Parch	Ticket	Fare	Cabin	Embarked
0	0	A/5 21171	7.2500	NaN	S
1	0	PC 17599	71.2833	C85	C
2	0	STON/O2. 3101282	7.9250	NaN	S
3	0	113803	53.1000	C123	S
4	0	373450	8.0500	NaN	S
..
886	0	211536	13.0000	NaN	S
887	0	112053	30.0000	B42	S
888	2	W./C. 6607	23.4500	NaN	S
889	0	111369	30.0000	C148	C
890	0	370376	7.7500	NaN	Q

[891 rows x 12 columns]

```
In [4]: #first 4 rows and 2 columns
df.iloc[0:4, 0:2]
```

```
Out[4]:
```

	PassengerId	Survived
0	1	0
1	2	1
2	3	1
3	4	1

```
In [5]: #all rows and 2 columns
df.iloc[0:,0:2]
```

```
Out[5]:
```

	PassengerId	Survived
0	1	0
1	2	1
2	3	1
3	4	1
4	5	0
...
886	887	0
887	888	1
888	889	0
889	890	1
890	891	0

891 rows x 2 columns



```
In [6]: #all rows and all columns
df.iloc[0:,]
```

Out[6]:

	PassengerId	Survived	Pclass	Name	gender	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	NaN	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

891 rows × 12 columns

```
In [7]: #all rows and 0th and 2nd column
df.iloc[0:,[0,2]]
```

Out[7]:

	PassengerId	Pclass
0	1	3
1	2	1
2	3	3
3	4	1
4	5	3
...
886	887	2
887	888	1
888	889	3
889	890	1
890	891	3

891 rows × 2 columns

```
In [8]: #any specific row and columns
df.iloc[[1,4,6],[0,1]]
```

Out[8]:

	PassengerId	Survived
1	2	1
4	5	0
6	7	0

In [10]: #all rows and Name and Age columns

```

In [10]: #all rows and Name and Age column
df.loc[:,["Name","Age"]]

Out[10]:

```

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

891 rows × 2 columns

```

In [11]: #correlation
cor=df.corr()
print(cor)

```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	\
PassengerId	1.000000	-0.005007	-0.035144	0.036847	-0.057527	-0.001652	
Survived	-0.005007	1.000000	-0.338481	-0.077221	-0.035322	0.081629	
Pclass	-0.035144	-0.338481	1.000000	-0.369226	0.083081	0.018443	
Age	0.036847	-0.077221	-0.369226	1.000000	-0.308247	-0.189119	
SibSp	-0.057527	-0.035322	0.083081	-0.308247	1.000000	0.414838	
Parch	-0.001652	0.081629	0.018443	-0.189119	0.414838	1.000000	
Fare	0.012658	0.257307	-0.549500	0.096067	0.159651	0.216225	

	Fare
PassengerId	0.012658
Survived	0.257307
Pclass	-0.549500
Age	0.096067
SibSp	0.159651
Parch	0.216225
Fare	1.000000

```
In [14]: #using filter we can sort the data
filt=df["gender"]=="female"
df.loc[filt,]
```

Out[14]:

PassengerId	Survived	Pclass	Name	gender	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	2	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
8	9	1	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C
...
880	881	1	Shelley, Mrs. William (Imanita Parrish Hall)	female	25.0	0	1	230433	26.0000	NaN	S
882	883	0	Dahlberg, Miss. Gerda Ulrika	female	22.0	0	0	7552	10.5167	NaN	S
885	886	0	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.1250	NaN	Q
887	888	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W/C. 6607	23.4500	NaN	S

314 rows × 12 columns

```
In [15]: #to calculate mean of Age column
agem=df["Age"].mean()
print(agem)
```

29.69911764705882

```
In [17]: #is used to update the mean at the missing places in the online dataframe
#or we can also use this command to fill the mean at missing places
#df.fillna(mean, inplace=True)
df["Age"]=df["Age"].fillna(agem)
df
```

Out[17]:

PassengerId	Survived	Pclass	Name	gender	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	Braund, Mr. Owen Harris	male	22.000000	1	0	A/5 21171	7.2500	NaN	S
1	2	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.000000	1	0	PC 17599	71.2833	C85	C
2	3	1	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.000000	1	0	113803	53.1000	C123	S
4	5	0	Allen, Mr. William Henry	male	35.000000	0	0	373450	8.0500	NaN	S
...
886	887	0	Montvila, Rev. Juozas	male	27.000000	0	0	211536	13.0000	NaN	S
887	888	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053	30.0000	B42	S
888	889	0	Johnston, Miss. Catherine Helen "Carrie"	female	29.699118	1	2	W/C. 6607	23.4500	NaN	S
889	890	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369	30.0000	C148	C
890	891	0	Dooley, Mr. Patrick	male	32.000000	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

```
In [18]: #is used to update the offline csv sheet
df.to_csv("F:\\SEM-5\\ML\\titanic-train - 1.csv")
```

Output:

Before updating the csv sheet:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Passenger	Survived	Pclass	Name	gender	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	
2	1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S	
3	2	1	1	Cumings, M	female	38	1	0	PC 17599	71.2833	C85	C	
4	3	1	3	Heikkinen, M	female	26	0	0	STON/O2.	7.925		S	
5	4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S	
6	5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S	
7	6	0	3	Moran, Mr	male		0	0	330877	8.4583		Q	
8	7	0	1	McCarthy, M	male	54	0	0	17463	51.8625	E46	S	
9	8	0	3	Palsson, M	male	2	3	1	349909	21.075		S	
10	9	1	3	Johnson, M	female	27	0	2	347742	11.1333		S	
11	10	1	2	Nasser, Mr	female	14	1	0	237736	30.0708		C	
12	11	1	3	Sandstrom	female	4	1	1	PP 9549	16.7	G6	S	
13	12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S	
14	13	0	3	Saunders, M	male	20	0	0	A/5. 2151	8.05		S	
15	14	0	3	Andersson	male	39	1	5	347082	31.275		S	
16	15	0	3	Vestrom, M	female	14	0	0	350406	7.8542		S	
17	16	1	2	Hewlett, M	female	55	0	0	248706	16		S	
18	17	0	3	Rice, Master	male	2	4	1	382652	29.125		Q	
19	18	1	2	Williams, M	male		0	0	244373	13		S	
20	19	0	3	Vander Planck	female	31	1	0	345763	18		S	
21	20	1	3	Masselmani	female		0	0	2649	7.225		C	
22	21	0	2	Fynney, M	male	35	0	0	239865	26		S	
23	22	1	2	Beesley, M	male	34	0	0	248698	13	D56	S	
24	23	1	3	McGowan	female	15	0	0	330923	8.0292		Q	
25	24	1	1	Sloper, Mr	male	28	0	0	113788	35.5	A6	S	
26	25	0	3	Palsson, M	female	8	3	1	349909	21.075		S	
27	26	1	3	Asplund, M	female	38	1	5	347077	31.3875		S	
28	27	0	3	Emir, Mr.	female		0	0	2631	7.225		C	
29	28	0	1	Fortune, M	male	19	3	2	19950	263	C23 C25 C	S	
30	29	1	3	O'Dwyer, M	female		0	0	330959	7.8792		Q	
31	30	0	3	Todoroff, M	male		0	0	349216	7.8958		S	
32	31	0	1	Urbach, M	male	40	0	0	PC 17601	27.7208		C	

After inserting the mean at missing value places in the csv sheet:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Passenger	Survived	Pclass	Name	gender	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked		
2	0	1	0	3 Braund, M	male	22	1	0	A/5 21171	7.25		S		
3	1	2	1	1 Cumings, M	female	38	1	0	PC 17599	71.2833	C85	C		
4	2	3	1	3 Heikkinen, M	female	26	0	0	STON/O2.	7.925		S		
5	3	4	1	1 Futrelle, M	female	35	1	0	113803	53.1	C123	S		
6	4	5	0	3 Allen, Mr.	male	35	0	0	373450	8.05		S		
7	5	6	0	3 Moran, Mr	male	29.69912	0	0	330877	8.4583		Q		
8	6	7	0	1 McCarthy, M	male	54	0	0	17463	51.8625	E46	S		
9	7	8	0	3 Palsson, M	male	2	3	1	349909	21.075		S		
10	8	9	1	3 Johnson, M	female	27	0	2	347742	11.1333		S		
11	9	10	1	2 Nasser, Mr	female	14	1	0	237736	30.0708		C		
12	10	11	1	3 Sandstrom	female	4	1	1	PP 9549	16.7	G6	S		
13	11	12	1	1 Bonnell, M	female	58	0	0	113783	26.55	C103	S		
14	12	13	0	3 Saunderson	male	20	0	0	A/5. 2151	8.05		S		
15	13	14	0	3 Andersson	male	39	1	5	347082	31.275		S		
16	14	15	0	3 Vestrom, M	female	14	0	0	350406	7.8542		S		
17	15	16	1	2 Hewlett, M	female	55	0	0	248706	16		S		
18	16	17	0	3 Rice, Mast	male	2	4	1	382652	29.125		Q		
19	17	18	1	2 Williams, M	male	29.69912	0	0	244373	13		S		
20	18	19	0	3 Vander Pla	female	31	1	0	345763	18		S		
21	19	20	1	3 Masselmani	female	29.69912	0	0	2649	7.225		C		
22	20	21	0	2 Fynney, M	male	35	0	0	239865	26		S		
23	21	22	1	2 Beesley, M	male	34	0	0	248698	13	D56	S		
24	22	23	1	3 McGowan	female	15	0	0	330923	8.0292		Q		
25	23	24	1	1 Sloper, Mr	male	28	0	0	113788	35.5	A6	S		
26	24	25	0	3 Palsson, M	female	8	3	1	349909	21.075		S		
27	25	26	1	3 Asplund, M	female	38	1	5	347077	31.3875		S		
28	26	27	0	3 Emir, Mr.	female	29.69912	0	0	2631	7.225		C		
29	27	28	0	1 Fortune, M	male	19	3	2	19950	263	C23 C25 C	S		
30	28	29	1	3 O'Dwyer, M	female	29.69912	0	0	330959	7.8792		Q		
31	29	30	0	3 Todoroff, M	male	29.69912	0	0	349216	7.8958		S		
32	30	31	0	1 Uruchurtu	male	40	0	0	PC 17601	27.7208		C		

Learning outcomes (What I have learnt):

1. Understanding of Exploratory Data analysis.
2. Able to analyze different datasets with the help of python and pandas library.
3. Learning about different library/packages of python.
4. Learning about the different methods, that are needed to analyze the given dataset.
5. Learning of different Machine Learning Functions

Evaluation Grid :

Sr. No.	Parameters	Marks Obtained	Maximum Marks
1.	Student Performance (Conduct of experiment) objectives/Outcomes.		12
2.	Viva Voce		10
3.	Submission of Work Sheet (Record)		8
	Total		30