Lab - 4 - Data Preprocessing

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

In [3]:

```
import pandas as pd
```

In [5]:

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

In [6]:

```
df.head()
```

Out[6]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cal
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	N
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	N
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	N

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

In [7]:

```
df
```

Out[7]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	C
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 [8]:

```
df.isnull()
```

Out[8]:

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

891 rows × 12 columns

In [9]:

```
df['Age'].isnull().sum
```

Out[9]:

```
<bound method NDFrame._add_numeric_operations.<locals>.sum of 0      F
alse
1      False
2      False
3      False
4      False
       ...
886    False
887    False
888     True
889    False
890    False
Name: Age, Length: 891, dtype: bool>
```

In [10]:

```
df[df['Age'].isnull()]
```

Out[10]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	C
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	
17	18	1	2	Williams, Mr. Charles Eugene	male	NaN	0	0	244373	13.0000	
19	20	1	3	Masselmani, Mrs. Fatima	female	NaN	0	0	2649	7.2250	
26	27	0	3	Emir, Mr. Farred Chehab	male	NaN	0	0	2631	7.2250	
28	29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female	NaN	0	0	330959	7.8792	
...	...	...	...	...	...	...	...	...	...	...	
859	860	0	3	Razi, Mr. Raihed	male	NaN	0	0	2629	7.2292	
863	864	0	3	Sage, Miss. Dorothy Edith "Dolly"	female	NaN	8	2	CA. 2343	69.5500	
868	869	0	3	van Melkebeke, Mr. Philemon	male	NaN	0	0	345777	9.5000	
878	879	0	3	Laleff, Mr. Kristo	male	NaN	0	0	349217	7.8958	
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	

177 rows × 12 columns

In [11]:

```
df2=df
```

In [12]:

```
df2.dropna()
```

Out[12]:

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

183 rows × 12 columns

In [14]:

```
df.fillna(0)
```

Out[14]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	C
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 [15]:

```
df
```

Out[15]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	C
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 [16]:

```
df2=df2.fillna({'Age':df["Age"].mean()})
```

In [17]:

```
df2.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          891 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
```

In [18]:

```
df2['Cabin']
```

Out[18]:

```
0       NaN
1       C85
2       NaN
3      C123
4       NaN
       ...
886     NaN
887     B42
888     NaN
889    C148
890     NaN
Name: Cabin, Length: 891, dtype: object
```

In [19]:

```
df2=df2.fillna({'Cabin':'C309'})
```

In [20]:

```
df2.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          891 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        891 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

In [21]:

```
df2['Embarked']
```

Out[21]:

```
0      S
1      C
2      S
3      S
4      S
      ..
886    S
887    S
888    S
889    C
890    Q
Name: Embarked, Length: 891, dtype: object
```

In [22]:

```
df2=df2.fillna({'Embarked':'S'})
```

In [23]:

```
df2.info()
```

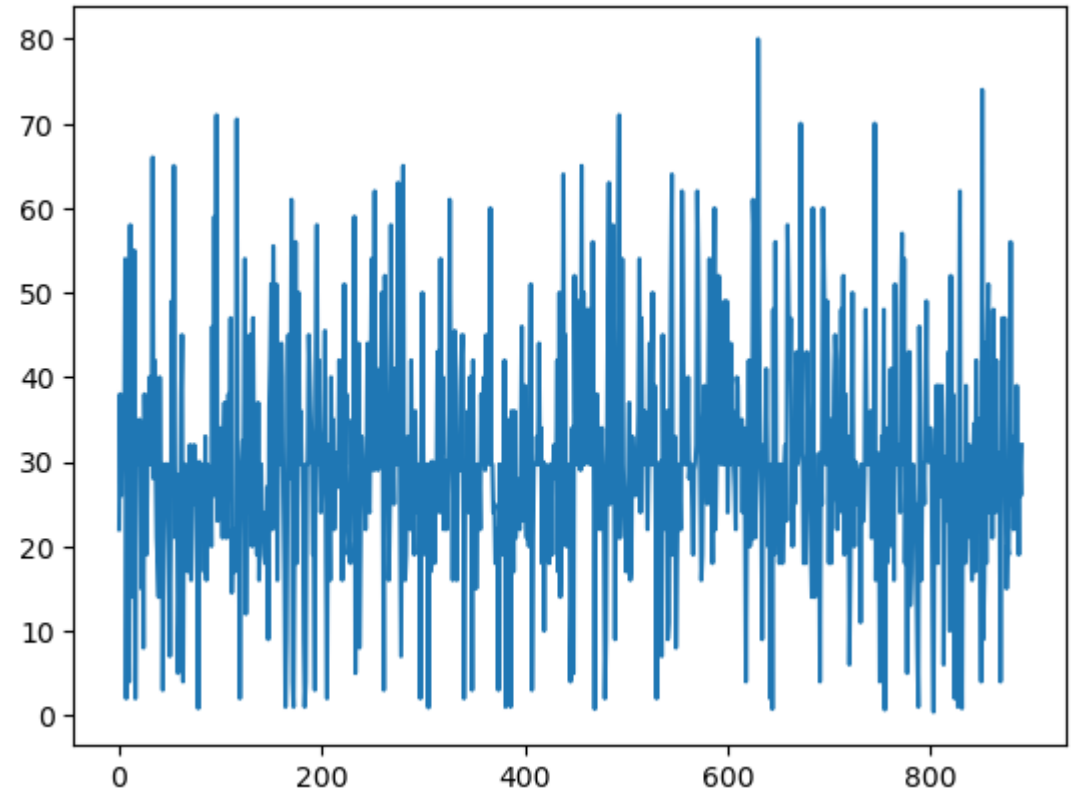
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null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

In [25]:

```
df2['Age'].plot()
```

Out[25]:

```
<AxesSubplot:>
```

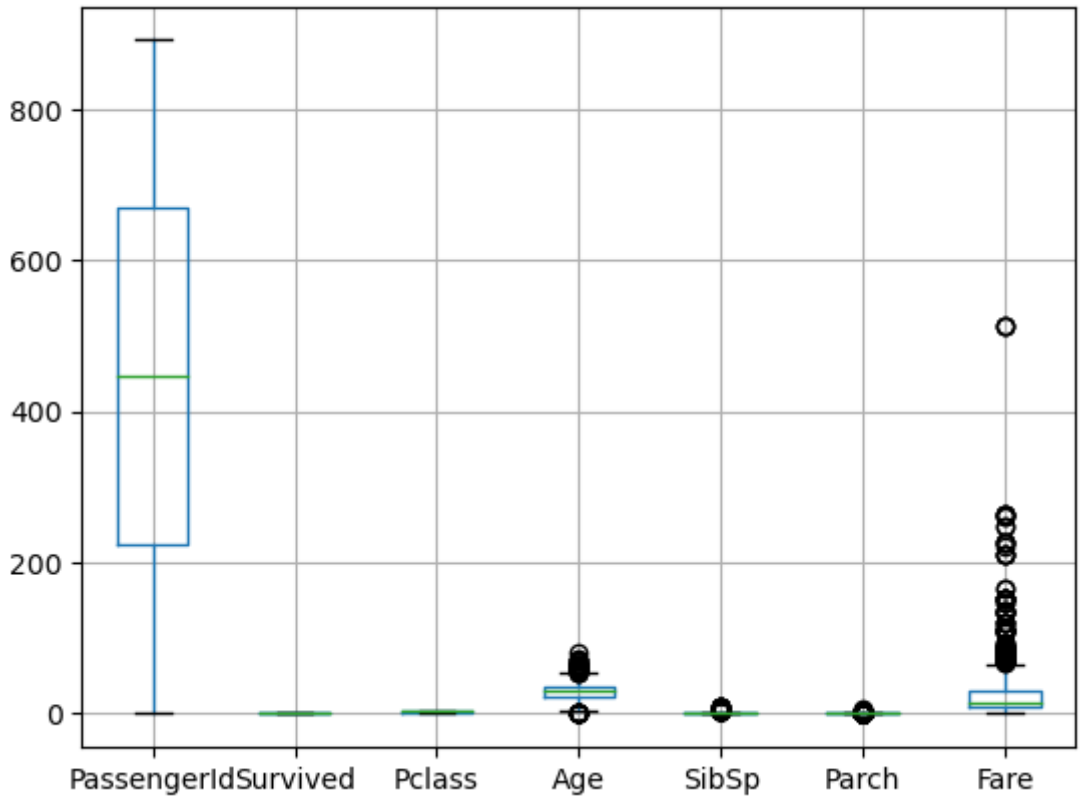


In [26]:

```
df2.boxplot()
```

Out[26]:

```
<AxesSubplot:>
```



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

In [27]:

```
df2["NewAge"] = df2['Age']+5
```

In [28]:

```
df2
```

Out[28]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
0	1	0	3	Braund, Mr. Owen Harris	male	22.000000	1	0	A/5 21171	7.25
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.000000	1	0	PC 17599	71.28
2	3	1	3	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282	7.92
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	8.05
...	...	...	...	...	...	...	...	...	...	
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	29.699118	1	2	W./C. 6607	23.45
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	7.75

891 rows × 13 columns

In [29]:

```
df2['NewAge1'] = ((df2['Age']-df2['Age'].min())/(df2['Age'].max()-df2['Age'].min()))
```

In [30]:

```
df2['Age'].max()
```

Out[30]:

```
80.0
```

In [31]:

```
n = df2['Age'].max()

d=0
while(n>0):
    n=n//10
    d=d+1
```

In [32]:

```
d
```

Out[32]:

```
2
```

In [33]:

```
df2["NewAge2"] = df2['Age']/(10**d)
```

In [34]:

```
df2["NewAge2"].max()
```

Out[34]:

```
0.8
```

In [35]:

```
df2
```

Out[35]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
0	1	0	3	Braund, Mr. Owen Harris	male	22.000000	1	0	A/5 21171	7.25
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.000000	1	0	PC 17599	71.28
2	3	1	3	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282	7.92
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	8.05
...	...	...	...	...	...	...	...	...	...	
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	29.699118	1	2	W./C. 6607	23.45
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	7.75

891 rows × 15 columns

In [36]:

```
df2[df2['Age'] == df2['Age'].max()]
```

Out[36]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	E
630	631	1	1	Barkworth, Mr. Algernon Henry Wilson	male	80.0	0	0	27042	30.0	A23	

In [37]:

```
df2['NewAge3'] = (df2['Age']-df2['Age'].mean())/df2['Age'].std()
```

In [38]:

```
df2
```

Out[38]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
0	1	0	3	Braund, Mr. Owen Harris	male	22.000000	1	0	A/5 21171	7.25
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.000000	1	0	PC 17599	71.28
2	3	1	3	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282	7.92
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.000000	1	0	113803	53.10
4	5	0	3	Allen, Mr. William Henry	male	35.000000	0	0	373450	8.05
...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536	13.00
887	888	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053	30.00
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.699118	1	2	W./C. 6607	23.45
889	890	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369	30.00
890	891	0	3	Dooley, Mr. Patrick	male	32.000000	0	0	370376	7.75

891 rows × 16 columns

In [39]:

```
df2['NewAge3'].max()
```

Out[39]:

```
3.868698926943564
```

In [40]:

```
df2[(df2['NewAge3'] > 3)]
```

Out[40]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Ca
96	97	0	1	Goldschmidt, Mr. George B	male	71.0	0	0	PC 17754	34.6542	
116	117	0	3	Connors, Mr. Patrick	male	70.5	0	0	370369	7.7500	C
493	494	0	1	Artagaveytia, Mr. Ramon	male	71.0	0	0	PC 17609	49.5042	C
630	631	1	1	Barkworth, Mr. Algernon Henry Wilson	male	80.0	0	0	27042	30.0000	A
672	673	0	2	Mitchell, Mr. Henry Michael	male	70.0	0	0	C.A. 24580	10.5000	C
745	746	0	1	Crosby, Capt. Edward Gifford	male	70.0	1	1	WE/P 5735	71.0000	B
851	852	0	3	Svensson, Mr. Johan	male	74.0	0	0	347060	7.7750	C

In [41]:

```
df2['NewAge3'].min()
```

Out[41]:

```
-2.2518907367915637
```

In [42]:

```
df2[(df2['NewAge3'] >= 2) | (df2['NewAge3'] <= -2)]
```

Out[42]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909	21.0750
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.00	0	0	113783	26.5500
16	17	0	3	Rice, Master. Eugene	male	2.00	4	1	382652	29.1250
33	34	0	2	Wheadon, Mr. Edward H	male	66.00	0	0	C.A. 24579	10.5000
43	44	1	2	Laroche, Miss. Simonne Marie Anne Andree	female	3.00	1	2	SC/Paris 2123	41.5792
...	...	...	...	...	...	...	...	...	...	...
827	828	1	2	Mallet, Master. Andre	male	1.00	0	2	S.C./PARIS 2079	37.0042
829	830	1	1	Stone, Mrs. George Nelson (Martha Evelyn)	female	62.00	0	0	113572	80.0000
831	832	1	2	Richards, Master. George Sibley	male	0.83	1	1	29106	18.7500
851	852	0	3	Svensson, Mr. Johan	male	74.00	0	0	347060	7.7750
879	880	1	1	Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)	female	56.00	0	1	11767	83.1583

69 rows × 16 columns

In [43]:

```
df2['NewFare'] = (df2['Fare']-df2['Fare'].mean())/df2['Fare'].std()
```

In [44]:

```
df2
```

Out[44]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
0	1	0	3	Braund, Mr. Owen Harris	male	22.000000	1	0	A/5 21171	7.25
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.000000	1	0	PC 17599	71.28
2	3	1	3	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282	7.92
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.000000	1	0	113803	53.10
4	5	0	3	Allen, Mr. William Henry	male	35.000000	0	0	373450	8.05
...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536	13.00
887	888	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053	30.00
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.699118	1	2	W./C. 6607	23.45
889	890	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369	30.00
890	891	0	3	Dooley, Mr. Patrick	male	32.000000	0	0	370376	7.75

891 rows × 17 columns

In [47]:

```
df2[(df2['NewFare'] >= 9) | (df2['NewFare'] <= -2)]
```

Out[47]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
258	259	1	1	Ward, Miss. Anna	female	35.0	0	0	PC 17755	512.3292	C30
679	680	1	1	Cardeza, Mr. Thomas Drake Martinez	male	36.0	0	1	PC 17755	512.3292	B5 B5 B5
737	738	1	1	Lesurer, Mr. Gustave J	male	35.0	0	0	PC 17755	512.3292	B10

In [50]:

```
df2[['Age','NewAge','NewAge1','NewAge2','NewAge3']].corr()
```

Out[50]:

	Age	NewAge	NewAge1	NewAge2	NewAge3
Age	1.0	1.0	1.0	1.0	1.0
NewAge	1.0	1.0	1.0	1.0	1.0
NewAge1	1.0	1.0	1.0	1.0	1.0
NewAge2	1.0	1.0	1.0	1.0	1.0
NewAge3	1.0	1.0	1.0	1.0	1.0