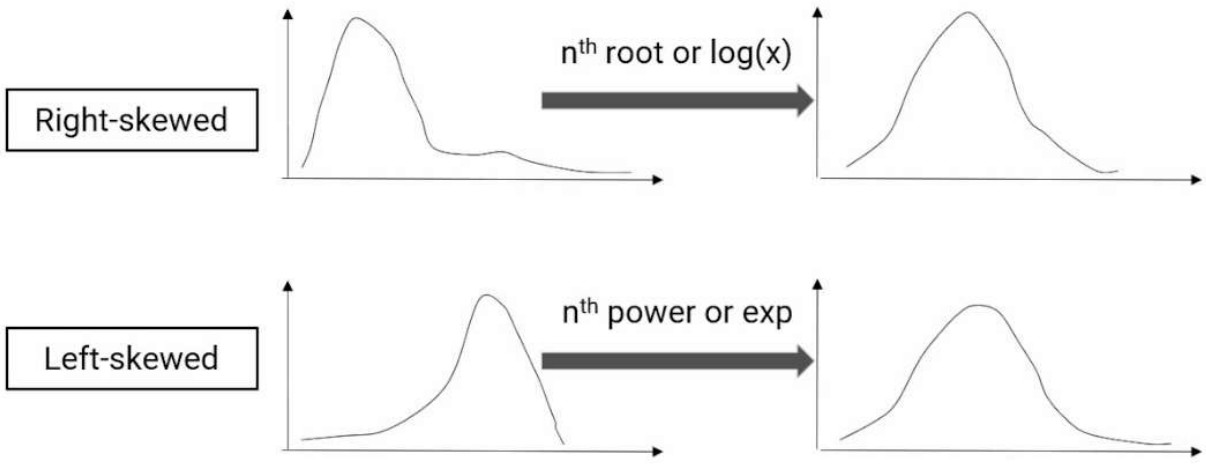
Feature Preprocessing: Transformation



In [1]:
```
#import packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

In [2]:
```
data=pd.read_csv('titanic.csv')
data.head()
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Na
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C8
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Na
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C12
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Na

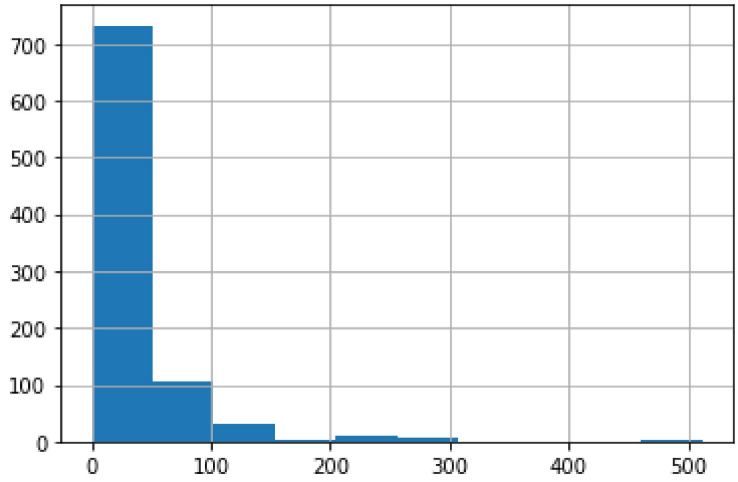
In [3]:
```
data=pd.read_csv('titanic.csv',usecols=['Fare','Age'])
data.head()
```

Out[3]:

	Age	Fare
0	22.0	7.2500
1	38.0	71.2833
2	26.0	7.9250
3	35.0	53.1000
4	35.0	8.0500

In [4]:
```
data['Fare'].hist()
```

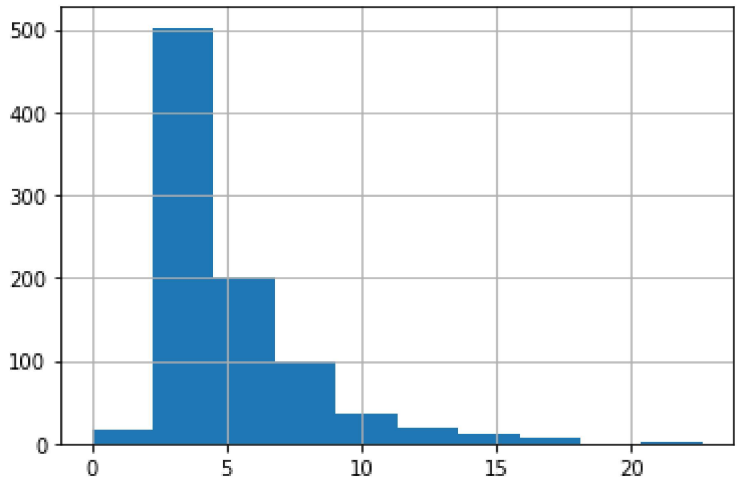
Out[4]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x24a3388ce80>
```



Root Transformation

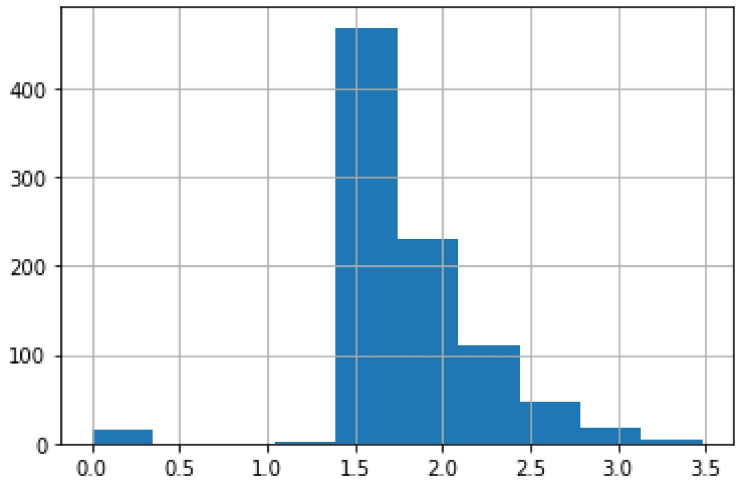
In [5]:
```
data['sqr_Fare']=data['Fare']**(1/2)
data['sqr_Fare'].hist()
```

Out[5]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x24a339c5cd0>
```



In [6]:
```
data['root_Fare']=data['Fare']**(1/5)
data['root_Fare'].hist()
```

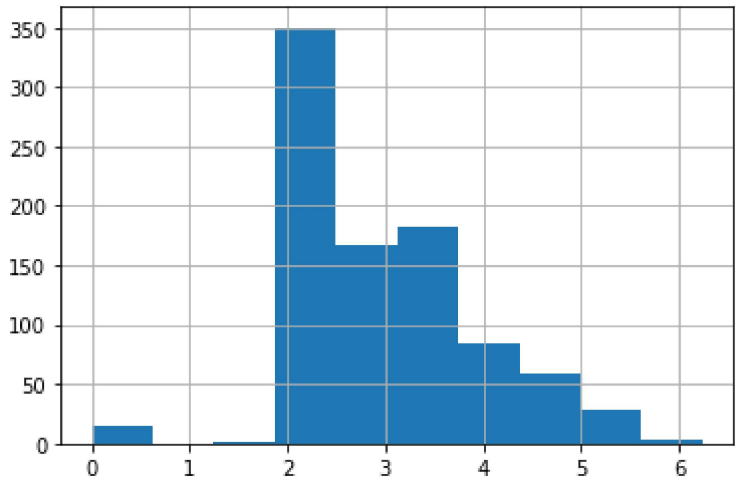
Out[6]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x24a33a46eb0>
```



Logarithmic Transformation

In [7]:
```
data['Log_Fare']=np.log(data['Fare']+1)
data['Log_Fare'].hist()
```

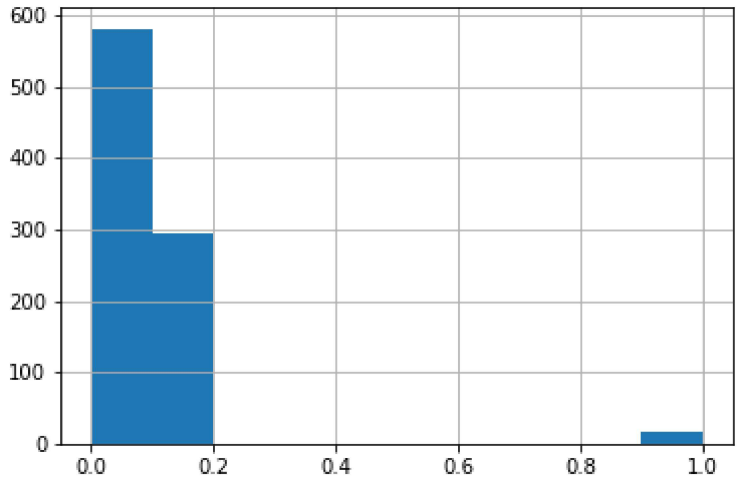
Out[7]: <matplotlib.axes._subplots.AxesSubplot at 0x24a33abc3a0>



Reciprocal transformation

In [8]:
```
data['Rec_Fare']=1/(data['Fare']+1)
data['Rec_Fare'].hist()
```

Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x24a33b40220>



BoxCox

The Box-Cox transformation is defined as:

$T(Y)=(Y\ \exp(\lambda)-1)/\lambda$

where Y is the response variable and λ is the transformation parameter. λ varies from -5 to 5. In the transformation, all values of λ are considered and the optimal value for a given variable is selected.

In [9]:
```
from scipy import stats
data['Fare_boxcox'], param = stats.boxcox(data.Fare+1) # you can vary the exponer
data['Fare_boxcox'].hist()
```

Out[9]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x24a35838370>
```