

Outlier Treatment

Loading the data

In [1]:

```
# importing required libraries
import pandas as pd
```

In [2]:

```
#Loading the data
data = pd.read_csv('titanic_train.csv')
data.head()
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	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

In [3]:

```
data.describe()
```

Out[3]:

	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

- Fare has some extreme values

In [4]:

```
data['Fare'].quantile(0.25), data['Fare'].quantile(0.75)
```

Out[4]:

(7.9104, 31.0)

In [5]:

```
IQR = data['Fare'].quantile(0.75) - data['Fare'].quantile(0.25)
IQR
```

Out[5]:

23.0896

In [6]:

```
Q1 = data['Fare'].quantile(0.25)
Q3 = data['Fare'].quantile(0.75)

whisker_1 = Q1 - (1.5*IQR)
whisker_2 = Q3 + (1.5*IQR)

whisker_1, whisker_2
```

Out[6]:

(-26.724, 65.6344)

Delete Outliers

In [7]:

```
data.loc[data['Fare']>66]
```

Out[7]:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	2	1	1Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
27	28	0	1Fortune, Mr. Charles Alexander	male	19.0	3	2	19950	263.0000	C23 C25 C27	S
31	32	1	1Spencer, Mrs. William Augustus (Marie Eugenie)	female	NaN	1	0	PC 17569	146.5208	B78	C
34	35	0	1Meyer, Mr. Edgar Joseph	male	28.0	1	0	PC 17604	82.1708	NaN	C
52	53	1	1Harper, Mrs. Henry Sleeper (Myna Haxtun)	female	49.0	1	0	PC 17572	76.7292	D33	C
...
846	847	0	3Sage, Mr. Douglas Bullen	male	NaN	8	2	CA. 2343	69.5500	NaN	S
849	850	1	1Goldenberg, Mrs. Samuel L (Edwiga Grabowska)	female	NaN	1	0	17453	89.1042	C92	C
856	857	1	1Wick, Mrs. George Dennick (Mary Hitchcock)	female	45.0	1	1	36928	164.8667	NaN	S
863	864	0	3Sage, Miss. Dorothy Edith "Dolly"	female	NaN	8	2	CA. 2343	69.5500	NaN	S
879	880	1	1Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)	female	56.0	0	1	11767	83.1583	C50	C

116 rows × 12 columns

In [8]:

```
data_new = data.loc[data['Fare']<66]  
data.shape, data_new.shape
```

Out[8]:

```
((891, 12), (775, 12))
```

Replacing Outliers

In [9]:

```
data['Fare'].loc[data['Fare']<7]
```

Out[9]:

```
129    6.9750  
143    6.7500  
179    0.0000  
202    6.4958  
263    0.0000  
271    0.0000  
277    0.0000  
302    0.0000  
326    6.2375  
371    6.4958  
378    4.0125  
411    6.8583  
413    0.0000  
466    0.0000  
481    0.0000  
597    0.0000  
633    0.0000  
654    6.7500  
674    0.0000  
732    0.0000  
804    6.9750  
806    0.0000  
815    0.0000  
818    6.4500  
822    0.0000  
825    6.9500  
843    6.4375  
872    5.0000  
Name: Fare, dtype: float64
```

In [10]:

```
data['Fare'].loc[data['Fare']<7] = Q1
```

C:\Users\pc\AppData\Local\Temp\ipykernel_17384\1197437457.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)

```
data['Fare'].loc[data['Fare']<7] = Q1
```

In [11]:

```
data['Fare'].loc[data['Fare']<7]
```

Out[11]:

```
Series([], Name: Fare, dtype: float64)
```

In [12]:

```
data['Fare'][872]
```

Out[12]:

7.9104

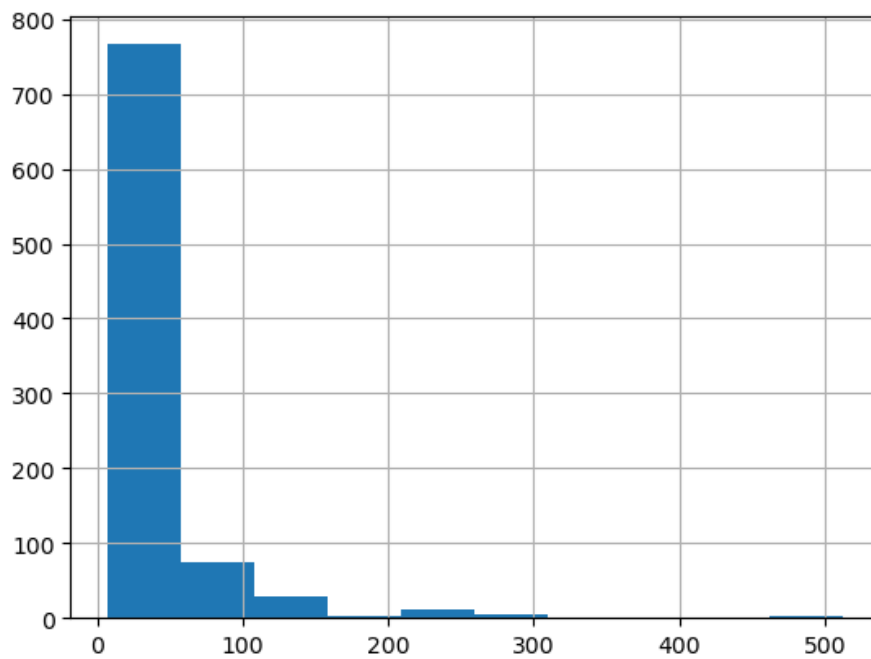
Transforming Variables

In [13]:

```
%matplotlib inline  
(data['Fare']).hist()
```

Out[13]:

<AxesSubplot:>



In [14]:

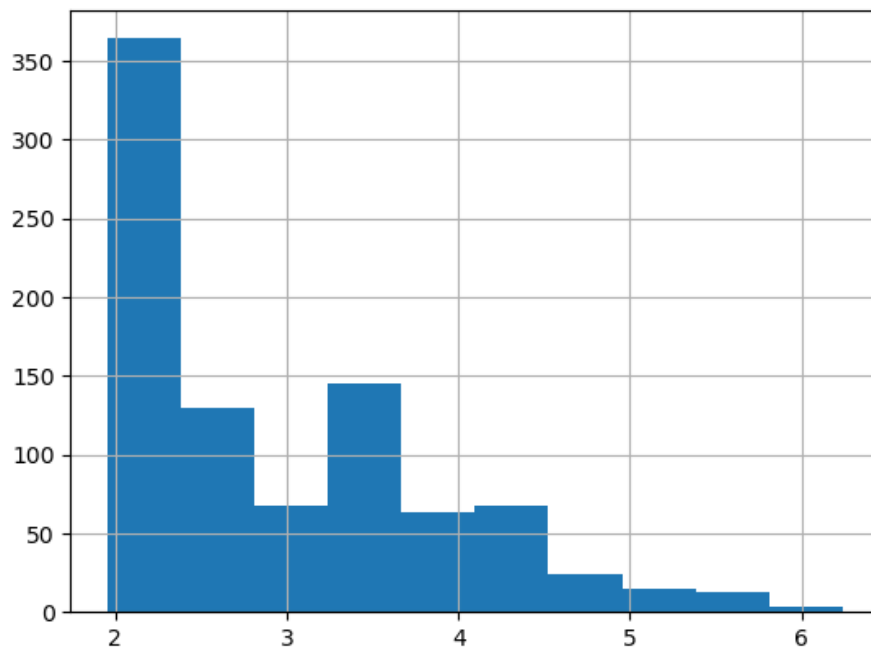
```
import numpy as np
```

In [15]:

```
(np.log(data['Fare'])).hist()
```

Out[15]:

<AxesSubplot:>



In []: