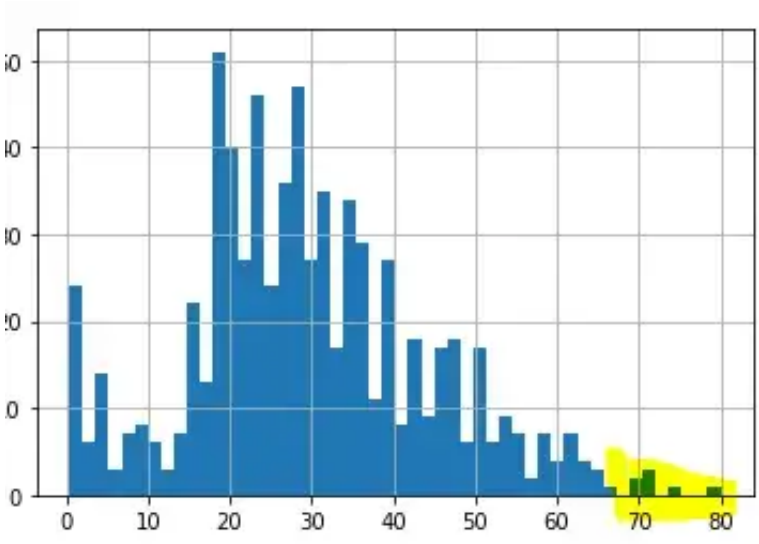
End of Distribution Imputation

End of Distribution is when we basically takes a value from end of the distribution(After third standard deviation)and replace nan with that value .



When we use it?

We use End of distribution in case missing completely at random(MCAR) process:

Advantages

Captures the importance of missing ness if there is any

Disadvantages

1. Distorts the original distribution of data
2. If NA is big, it will mask outliers
3. If NA is small, the replaced NA will be considered as outlier
4. Arbitrary value imputation: It means imputing missing values with an arbitrary value

Example

In [65]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import missingno as msno
%matplotlib inline

data=pd.read_csv('test.csv')
```

In [66]:
```
data
```

Out[66]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	C
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	
...	...	...	...	...	...	...	...	...	...	
413	1305	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	8.0500	
414	1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C
415	1307	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	
416	1308	3	Ware, Mr. Frederick	male	NaN	0	0	359309	8.0500	
417	1309	3	Peter, Master. Michael J	male	NaN	1	1	2668	22.3583	

418 rows × 11 columns

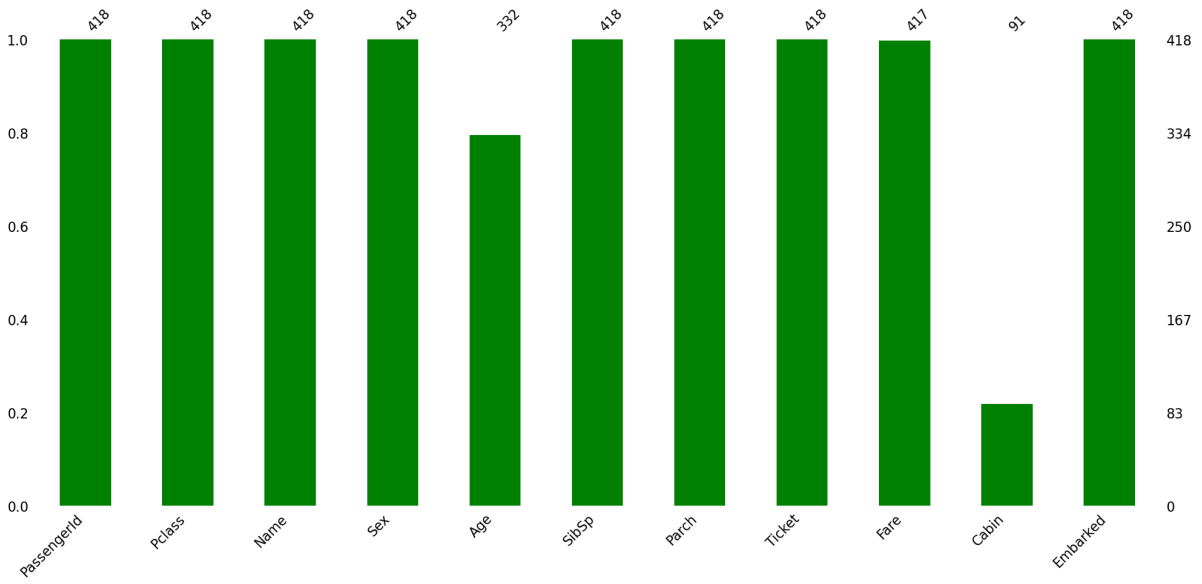
In [67]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  418 non-null    int64  
 1   Pclass       418 non-null    int64  
 2   Name         418 non-null    object 
 3   Sex          418 non-null    object 
 4   Age          332 non-null    float64
 5   SibSp        418 non-null    int64  
 6   Parch        418 non-null    int64  
 7   Ticket       418 non-null    object 
 8   Fare         417 non-null    float64
 9   Cabin        91 non-null     object 
 10  Embarked     418 non-null    object 
dtypes: float64(2), int64(4), object(5)
memory usage: 36.0+ KB
```

In [68]:
```
data.isna().sum()
```

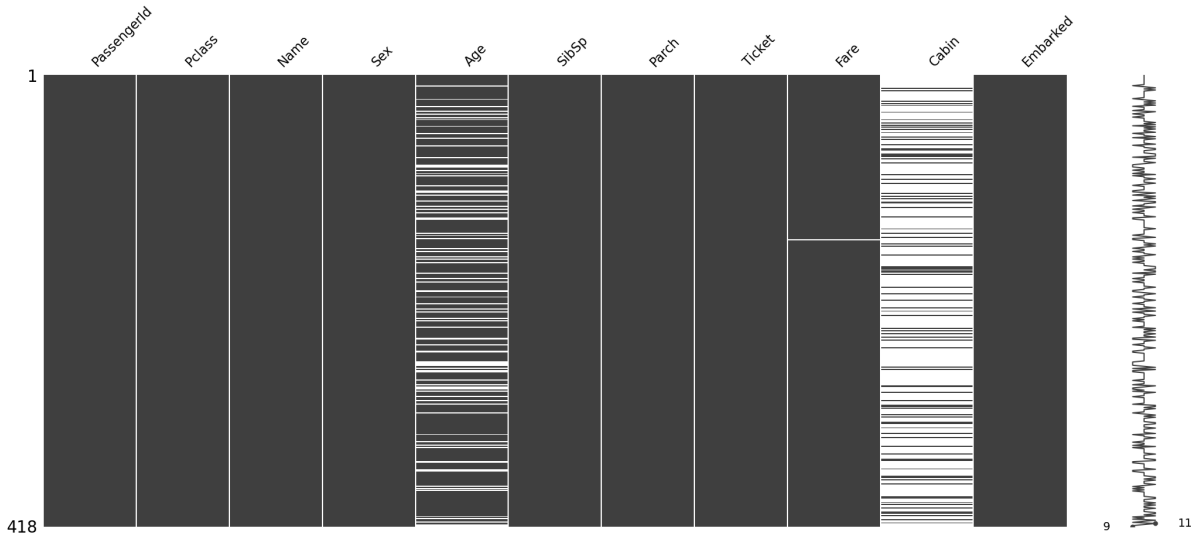
Out[68]:
```
PassengerId      0
Pclass           0
Name             0
Sex              0
Age             86
SibSp            0
Parch            0
Ticket           0
Fare             1
Cabin          327
Embarked         0
dtype: int64
```

In [71]:
```
chart=msno.bar(data,color='g')
```



In [75]:
```
msno.matrix(data)
```

Out[75]: (<AxesSubplot:>,)



Handing with missing value by using End of Distribution Imputation:

In [56]:
```
extreme=data.Age.mean()+3*data.Age.std()
median=data.Age.median()
```

In [57]:
```
def impute_nan (data,variable,median,extreme):
    data[variable+'_end_distribution']=data[variable].fillna(extrem
    data[variable].fillna(median,inplace=True)
```

In [58]:
```
impute_nan(data,'Age',median,extreme)
```

In [59]:
```
data
```

Out[59]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Ca
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	
...	...	...	...	...	...	...	...	...	...	
413	1305	3	Spector, Mr. Woolf	male	27.0	0	0	A.5. 3236	8.0500	
414	1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C
415	1307	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	
416	1308	3	Ware, Mr. Frederick	male	27.0	0	0	359309	8.0500	
417	1309	3	Peter, Master. Michael J	male	27.0	1	1	2668	22.3583	

418 rows × 12 columns

In [62]:
```
data.isna().sum()
```

Out[62]:
```
PassengerId              0
Pclass                   0
Name                     0
Sex                      0
Age                      0
SibSp                    0
Parch                    0
Ticket                   0
Fare                     1
Cabin                  327
Embarked                 0
Age_end_distribution     0
dtype: int64
```

In [ ]: