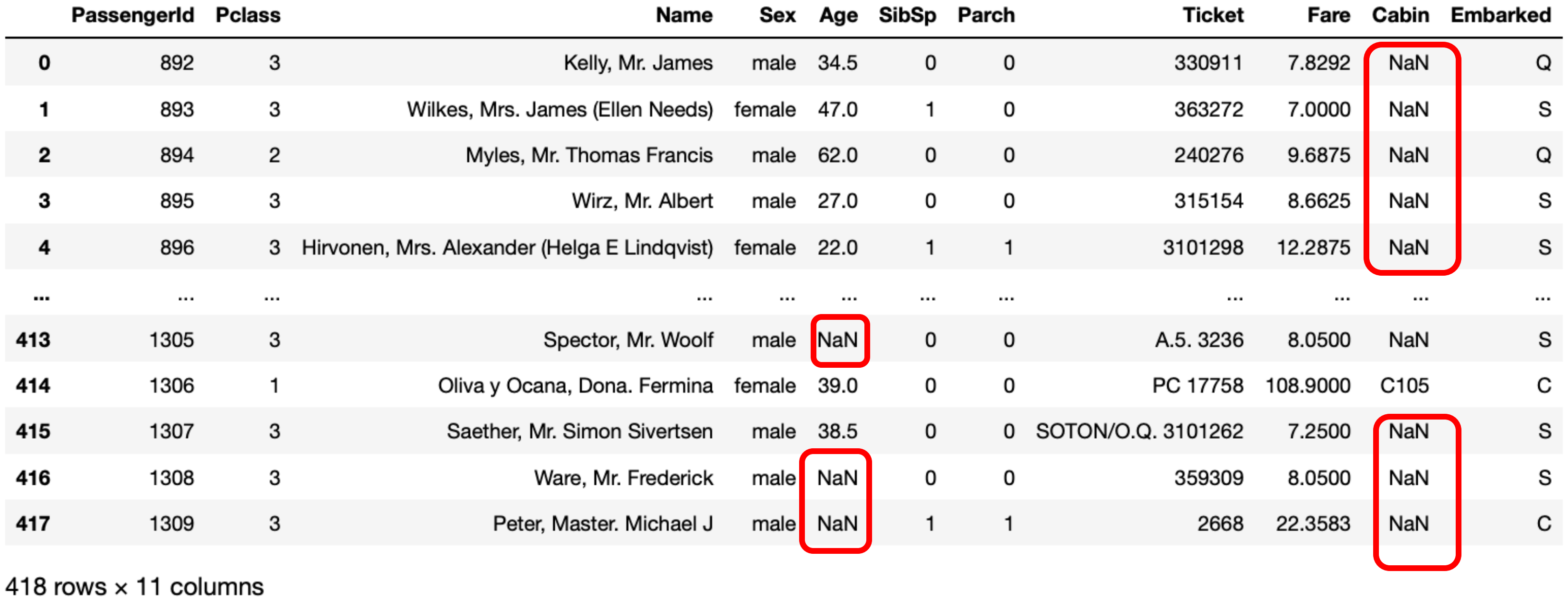
End of distribution

Missing Value

Prepare by Roatny NUON

What is missing Data?

- Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset.



	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...
413	1305	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	8.0500	NaN	S
414	1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C105	C
415	1307	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	NaN	S
416	1308	3	Ware, Mr. Frederick	male	NaN	0	0	359309	8.0500	NaN	S
417	1309	3	Peter, Master. Michael J	male	NaN	1	1	2668	22.3583	NaN	C

418 rows × 11 columns

Type of missing value?

There are 3 types of missing value:

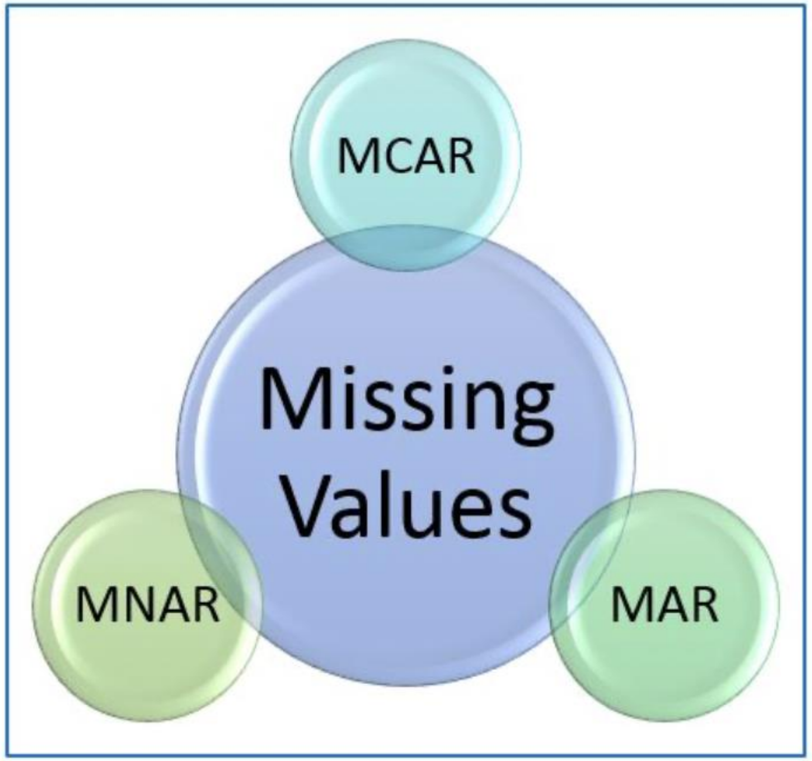


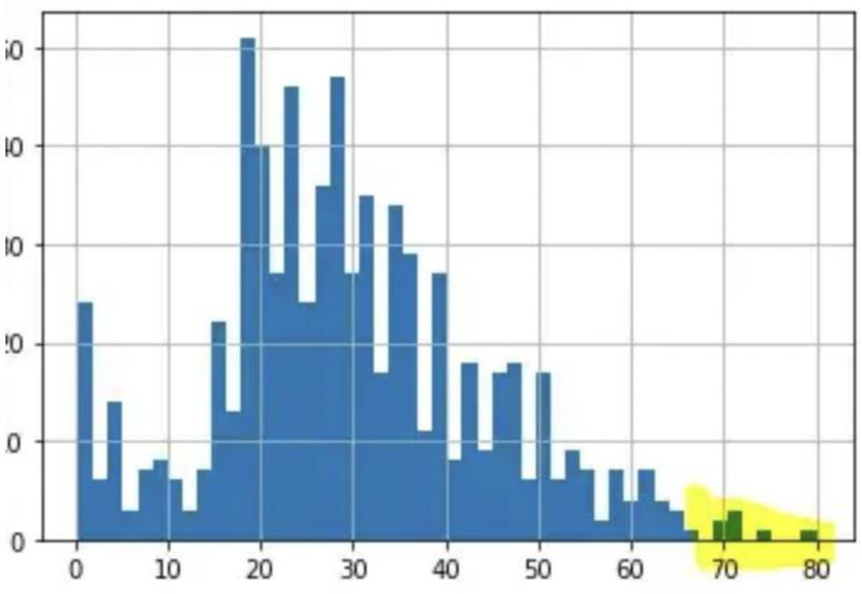
Figure 1 - Different Types of Missing Values in Datasets

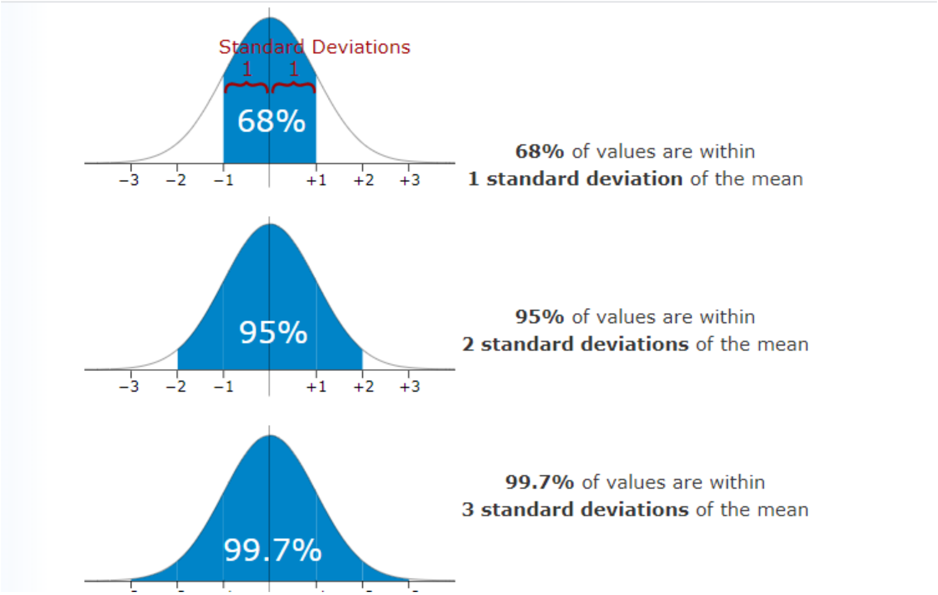
Missing Completely At Random (MCAR)

- In MCAR, the probability of data being missing is the same for all the observations.
- In the case of MCAR, the data could be missing due to human error, some system/equipment failure, loss of sample, or some unsatisfactory technicalities while recording the values.

What is End of Distribution?

When we basically takes a value from end of the distribution (After third standard deviation)and replace nan with that value





68% of values are within
1 standard deviation of the mean

95% of values are within
2 standard deviations of the mean

99.7% of values are within
3 standard deviations of the mean

When we use End of Distribution?

It is also used in case of missing completely at random(MCAR), <5% missing value.

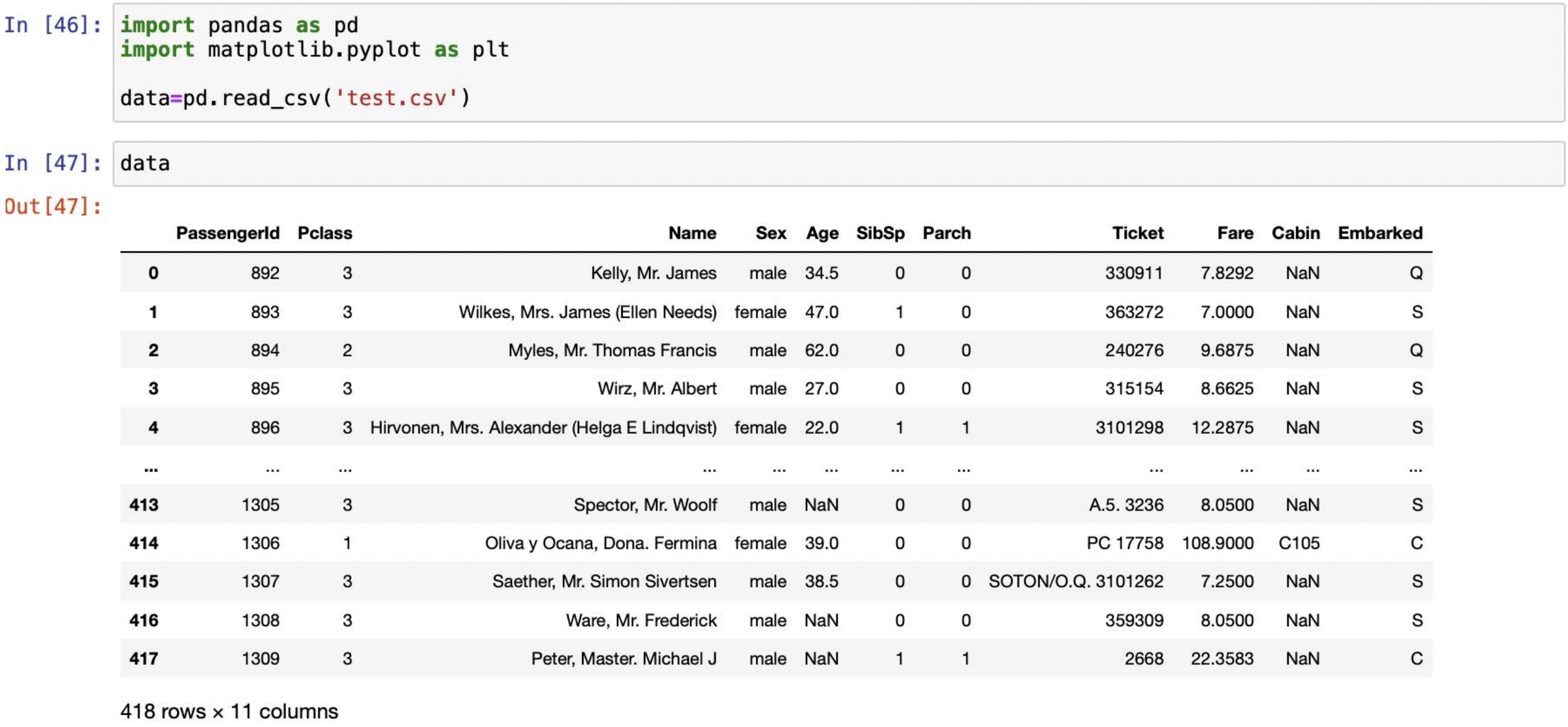
Advantages:

- Captures the importance of missing ness if there is any

Disadvantages:

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

Example:



```
In [46]: import pandas as pd
         import matplotlib.pyplot as plt

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

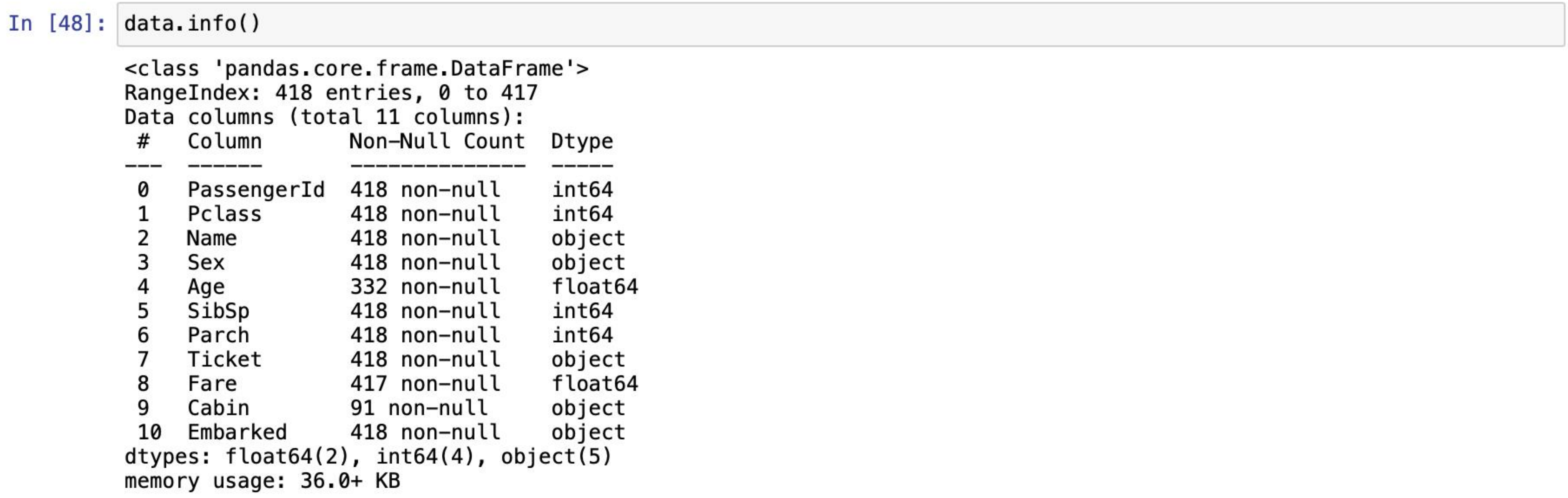
```
In [47]: data
```

Out[47]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...
413	1305	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	8.0500	NaN	S
414	1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C105	C
415	1307	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	NaN	S
416	1308	3	Ware, Mr. Frederick	male	NaN	0	0	359309	8.0500	NaN	S
417	1309	3	Peter, Master. Michael J	male	NaN	1	1	2668	22.3583	NaN	C

418 rows × 11 columns

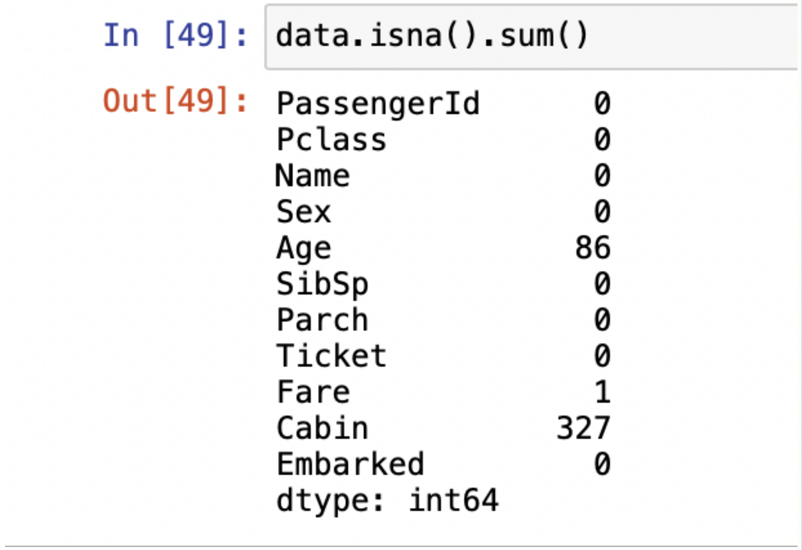
Check its info



```
In [48]: data.info()
```

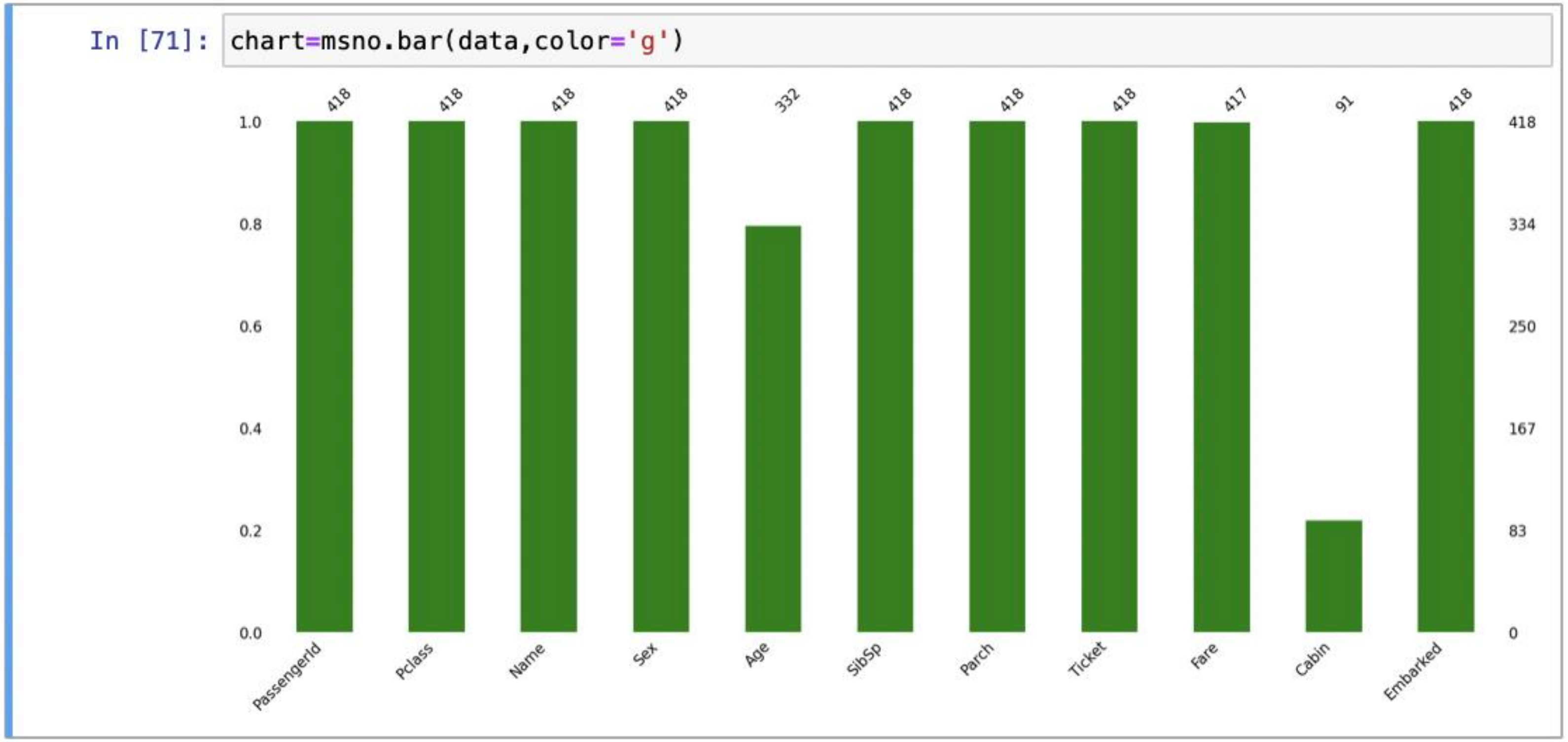
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  418 non-null    int64  
 1   Pclass       418 non-null    int64  
 2   Name         418 non-null    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
```

Check missing value



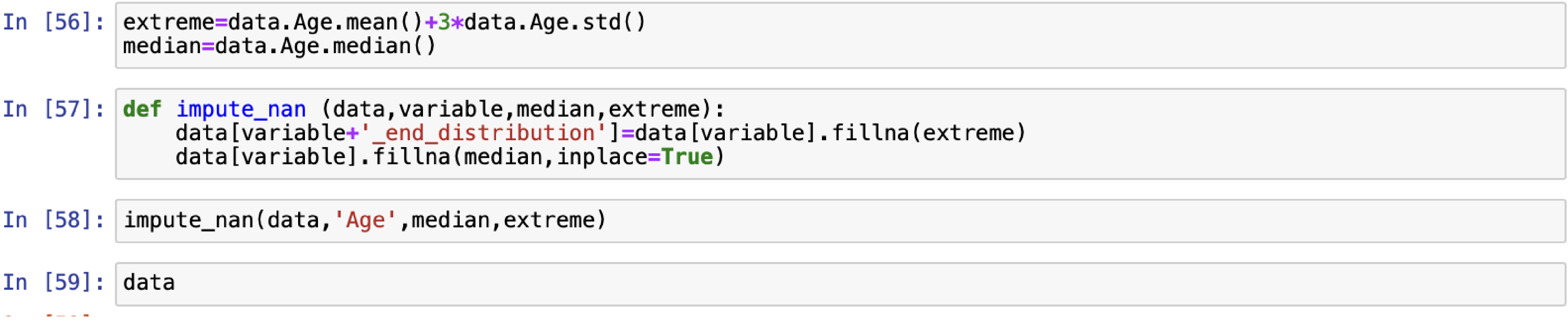
```
In [49]: data.isna().sum()

Out[49]: 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')
```

Handing with missing value by End of Distribution Imputation:

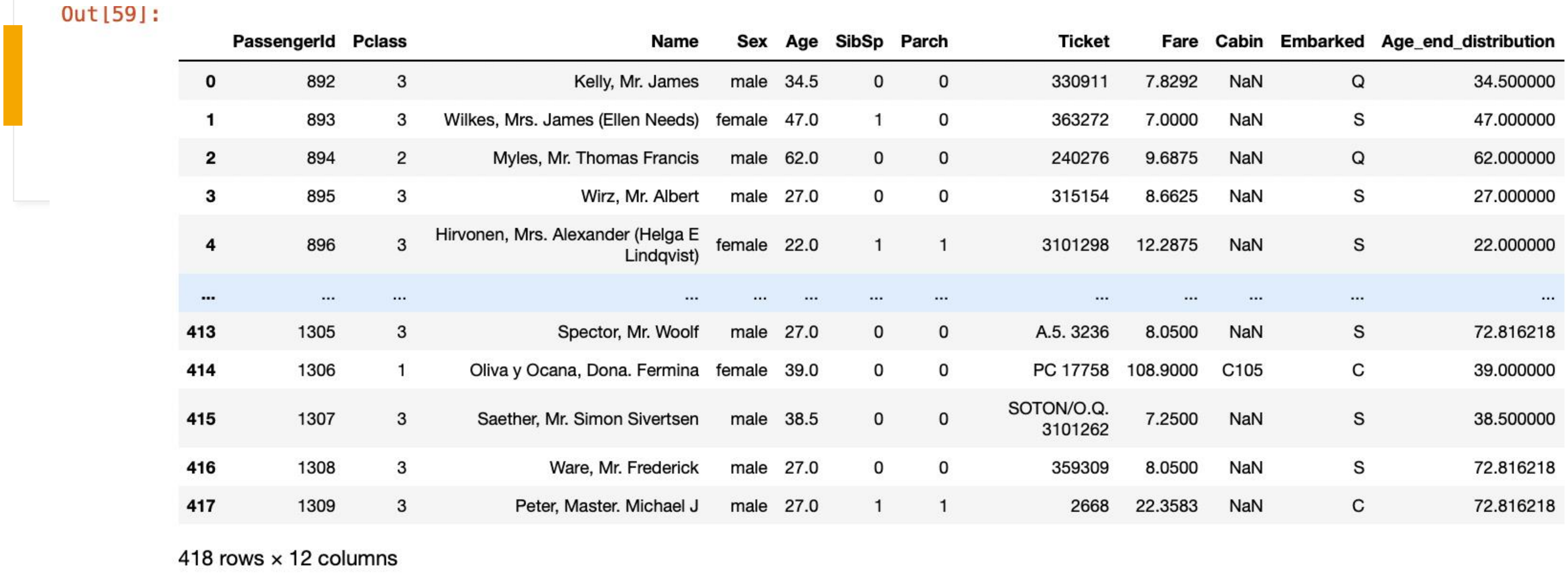


```
In [56]: extreme=data.Age.mean()+3*data.Age.std()
         median=data.Age.median()

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

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

In [59]: data
```

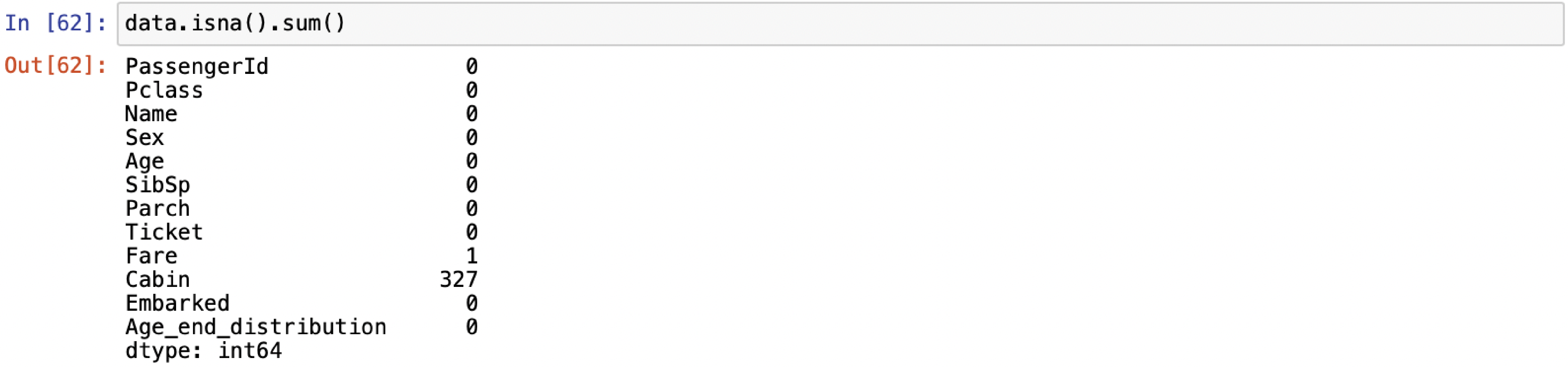


Out[59]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Age_end_distribution
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q	34.500000
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S	47.000000
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q	62.000000
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S	27.000000
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S	22.000000
...	...	...	...	...	...	...	...	...	...	...	...	...
413	1305	3	Spector, Mr. Woolf	male	27.0	0	0	A.5. 3236	8.0500	NaN	S	72.816218
414	1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C105	C	39.000000
415	1307	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	NaN	S	38.500000
416	1308	3	Ware, Mr. Frederick	male	27.0	0	0	359309	8.0500	NaN	S	72.816218
417	1309	3	Peter, Master. Michael J	male	27.0	1	1	2668	22.3583	NaN	C	72.816218

418 rows × 12 columns

Check if it works, Good to GO!



```
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
```