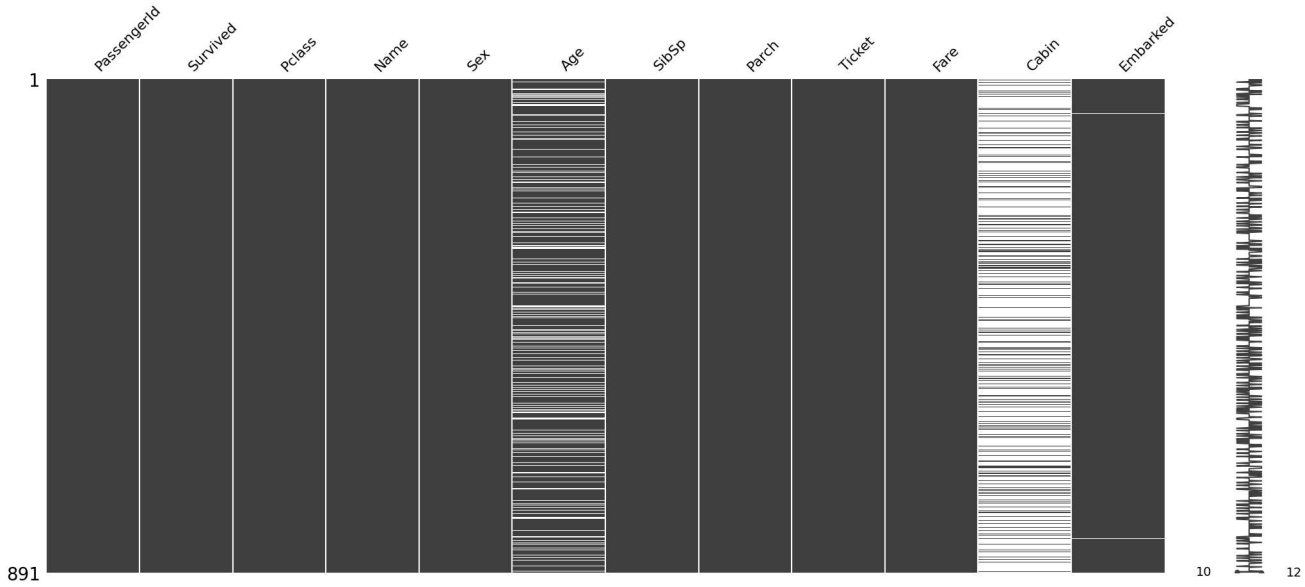
7 Ways to Handle Missing Values

```
from sklearn.linear_model import LinearRegression
import pandas as pd
import numpy as np
import missingno as msno
from feature_engine.imputation import MeanMedianImputer
```

```
data = pd.read_csv("train.csv")
```

```
msno.matrix(data)
```

```
<Axes: >
```




```
# print(data.isnull().sum())
# print(data.shape)
```

```
# data.dropna(inplace=True)
# print(data.isnull().sum())
# print(data.shape)
```

```
# data["Age"][:20]
```

```
# data["Age"] = data["Age"].replace(np.NaN, data["Age"].mean())
# print(data["Age"][:20])
```

```
data.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
```

```
Name            0
Sex             0
Age           177
SibSp           0
Parch           0
Ticket          0
Fare            0
Cabin         687
Embarked        2
dtype: int64
```

```
data["Cabin"] = data["Cabin"].fillna('U')
```

```
data.isnull().sum()
```

```
PassengerId     0
Survived        0
Pclass          0
Name            0
Sex             0
Age           177
SibSp           0
Parch           0
Ticket          0
Fare            0
Cabin           0
Embarked        2
dtype: int64
```

```
print(data["Age"][:20])
```

```
0     22.0
1     38.0
2     26.0
3     35.0
4     35.0
5      NaN
6     54.0
7      2.0
8     27.0
9     14.0
10     4.0
11    58.0
12    20.0
13    39.0
14    14.0
15    55.0
16     2.0
17     NaN
18    31.0
19     NaN
Name: Age, dtype: float64
```

```
data["Age"] = data["Age"].fillna(method='ffill')
```

```
print(data["Age"][:20])
```

```
0     22.0
1     38.0
2     26.0
3     35.0
4     35.0
5     35.0
6     54.0
7      2.0
8     27.0
9     14.0
10     4.0
11    58.0
12    20.0
13    39.0
14    14.0
15    55.0
16     2.0
17     2.0
18    31.0
19    31.0
Name: Age, dtype: float64
```

```
data["Age"] = data["Age"].interpolate(method='linear', limit_direction='forward', axis=0)
```

```
data.isnull().sum()
```

```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin          0
Embarked       2
dtype: int64
```

```
# data["Sex"] = [1 if x=="male" else 0 for x in data["Sex"]]
# print(data)
```

```
# test_data = data[data["Age"].isnull()]
# print(test_data)
# data.dropna(inplace=True)
```

```
# y_train = data["Age"]
# print(y_train)
# X_train = data.drop("Age", axis=1)
# X_test = test_data.drop("Age", axis=1)
```

```
# model = LinearRegression()
# model.fit(X_train, y_train)
```

```
# y_pred = model.predict(X_test)
# print(y_pred)
```

```
pip install feature-engine
```

```
Requirement already satisfied: feature-engine in /usr/local/lib/python3.10/dist-packa
Requirement already satisfied: numpy>=1.18.2 in /usr/local/lib/python3.10/dist-packag
Requirement already satisfied: pandas>=1.0.3 in /usr/local/lib/python3.10/dist-packag
Requirement already satisfied: scikit-learn>=1.0.0 in /usr/local/lib/python3.10/dist-
Requirement already satisfied: scipy>=1.4.1 in /usr/local/lib/python3.10/dist-package
Requirement already satisfied: statsmodels>=0.11.1 in /usr/local/lib/python3.10/di
Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/lib/python3.10
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-package
Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.10/dist-packag
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist
Requirement already satisfied: patsy>=0.5.2 in /usr/local/lib/python3.10/dist-package
Requirement already satisfied: packaging>=21.3 in /usr/local/lib/python3.10/dist-pack
Requirement already satisfied: six in /usr/local/lib/python3.10/dist-packages (from p
```

```
import pandas as pd
import numpy as np
from feature_engine.imputation import MeanMedianImputer
```

```
X = pd.DataFrame(dict(
       x1 = [np.nan,1,1,0,np.nan],
       x2 = ["a", np.nan, "b", np.nan, "a"],
       ))
mmi = MeanMedianImputer(imputation_method='median')
mmi.fit(X)
mmi.transform(X)
```

	x1	x2
0	1.0	a
1	1.0	NaN
2	1.0	b
3	0.0	NaN
4	1.0	a

```
mmi = MeanMedianImputer(imputation_method='median', variables=['Age'])
data['Age'] = mmi.fit_transform(data[['Age']])
print(data.head())
```

```
   PassengerId  Survived  Pclass  \
0            1         0       3
```

```
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500     U        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250     U        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500     U        S
```