```
#data analysis libraries
import numpy as np
import pandas as pd

#visualization libraries
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

#ignore warnings
import warnings
warnings.filterwarnings('ignore')
```

Start coding or generate with AI.

Double-click (or enter) to edit

```
data=pd.read_csv('/content/Titanic-Dataset.csv')
data
```

	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
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C

```
data.describe()
```

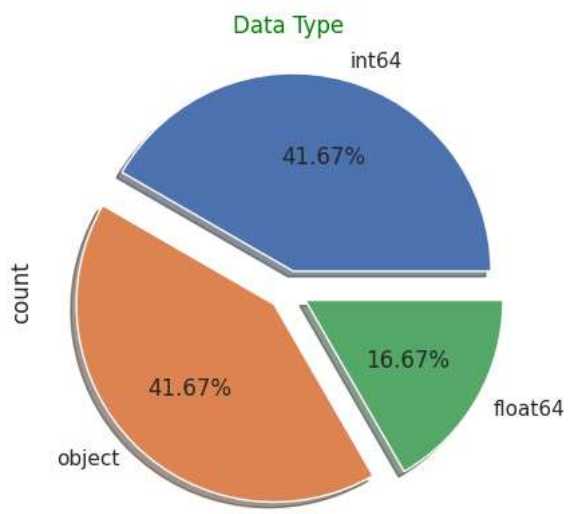
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	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

```
data.info()
```

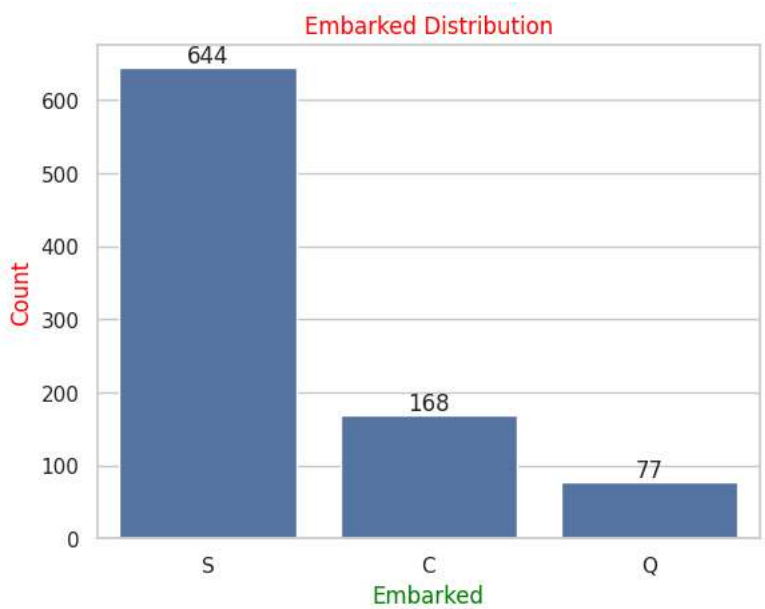
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
```

```
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

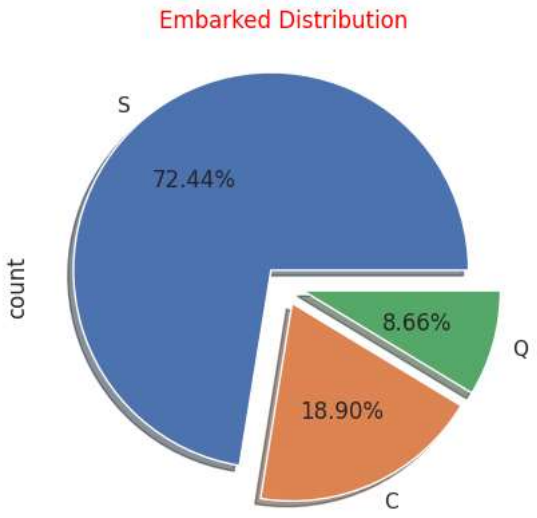
```
data.dtypes.value_counts().plot.pie(explode=[0.1, 0.1, 0.1],
                                    autopct='%1.2f%%',
                                    shadow=True)
plt.title('Data Type',
          color='Green',
          loc='center',
          );
```



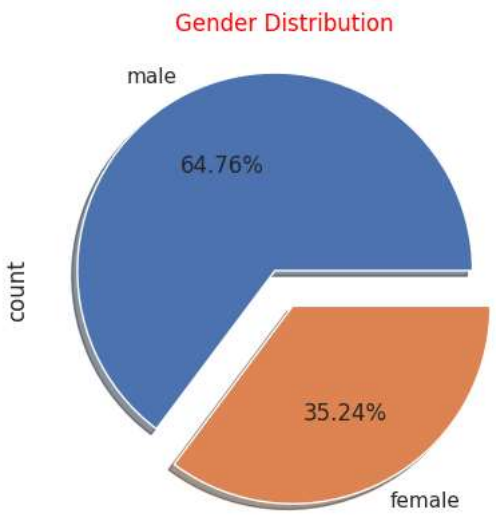

```
ax = sns.set(style="whitegrid")
ax = sns.countplot(data=data,x='Embarked');
ax.bar_label(ax.containers[0])

plt.title('Embarked Distribution',color='Red',loc='center');
plt.xlabel('Embarked',color='Green',loc='center')
plt.ylabel('Count',color='Red',loc='center');
```



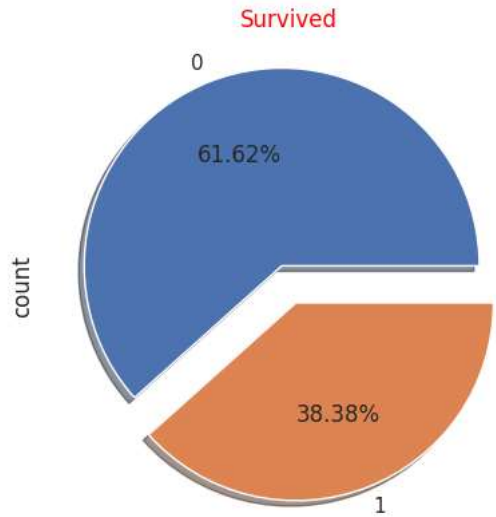
```
data['Embarked'].value_counts().plot.pie(explode=[0.1, 0.1, 0.1],
                                         autopct='%1.2f%%',
                                         shadow=True)
plt.title('Embarked Distribution',color='Red',loc='center');
```



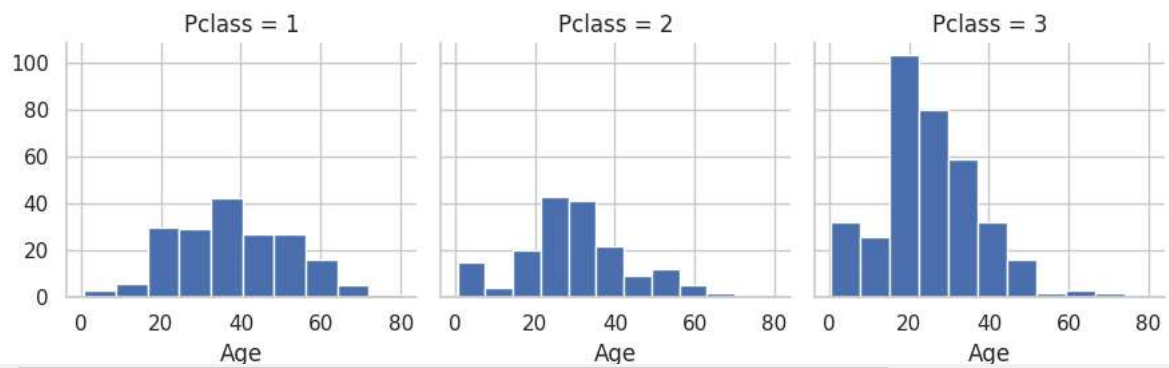

```
data['Sex'].value_counts().plot.pie(explode=[0.1, 0.1],
                                    autopct='%1.2f%%',
                                    shadow=True)
plt.title('Gender Distribution',color='Red',loc='center');
```




```
data['Survived'].value_counts().plot.pie(explode=[0.1, 0.1],
                                         autopct='%1.2f%%',
                                         shadow=True)
plt.title('Survived',color='Red',loc='center');
```




```
g = sns.FacetGrid(data, col="Pclass")
g = g.map(plt.hist, "Age")
```

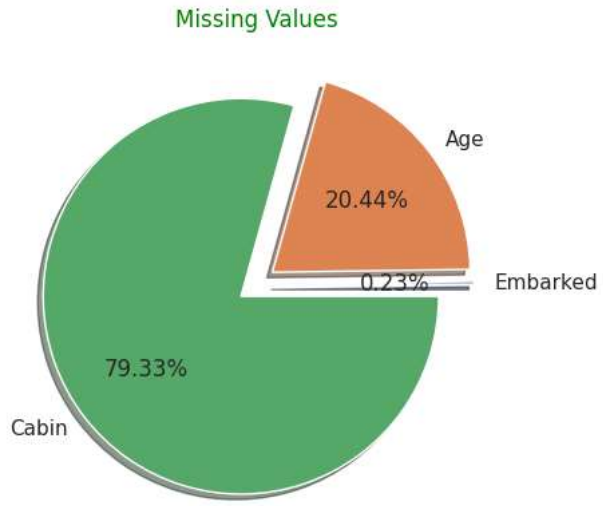


Start coding or generate with AI.

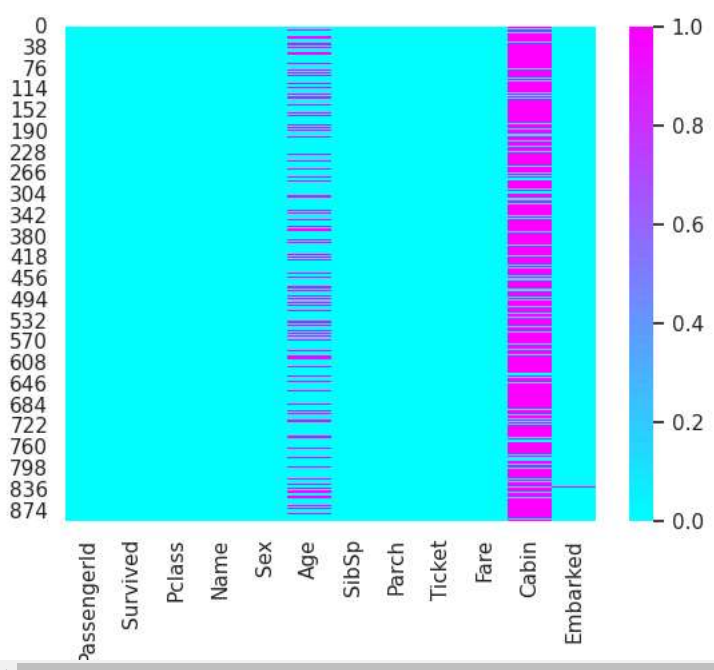
```
def missing_value (df):
    missing_Number = df.isnull().sum().sort_values(ascending=False)[df.isnull().sum().sort_values(ascending=False) !=0]
    missing_percent=round((df.isnull().sum()/df.isnull().count())*100,2)[round((df.isnull().sum()/df.isnull().count())*100,2) !=0]
    missing = pd.concat([missing_Number,missing_percent],axis=1,keys=['Missing Number','Missing Percentage'])
    return missing

# Apply styling with precision setting
styled_df = missing_value(data).style.background_gradient(cmap='coolwarm')
styled_df.format("{:.2f}", subset=['Missing Percentage']) # Format 'Missing Percentage' column with 2 decimal places
styled_df
```

	Missing Number	Missing Percentage
Cabin	687	77.10
Age	177	19.87
Embarked	2	0.22

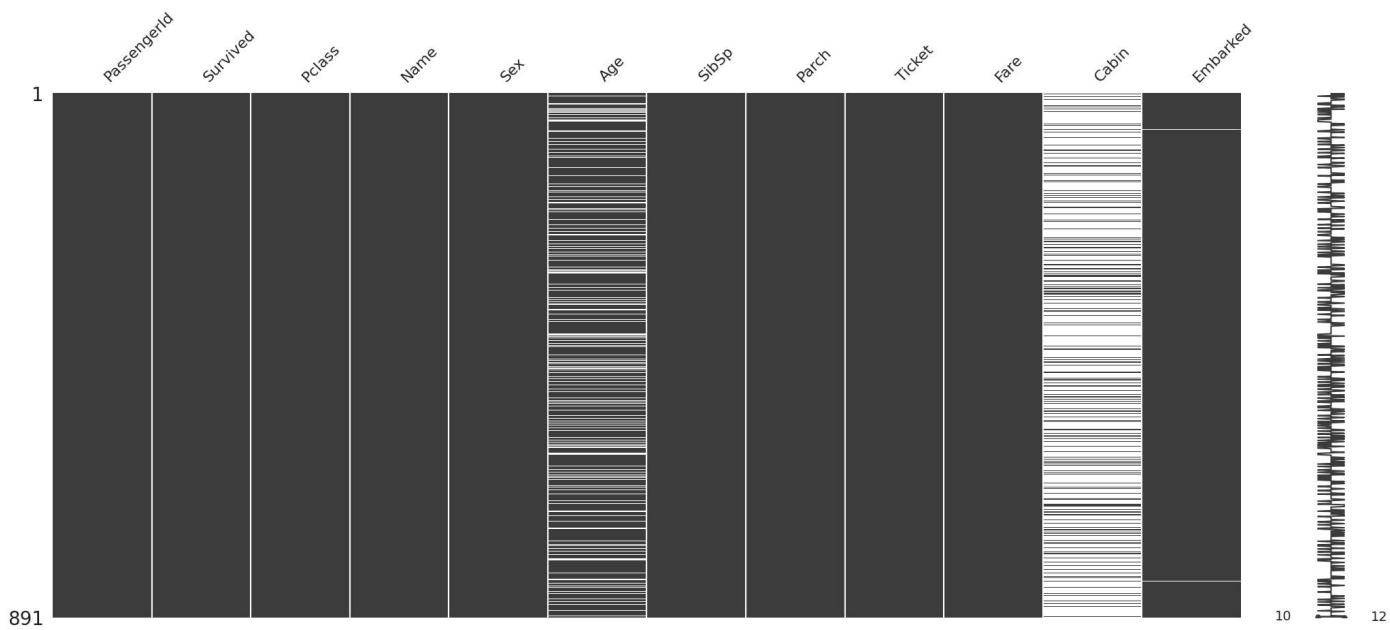
```
missing_values = data.isnull().sum()
missing_values = missing_values[missing_values > 0]
missing_values.sort_values(inplace=True)
missing_values.plot.pie(explode=[0.1, 0.1, 0.1],
                        autopct='%1.2f%%',
                        shadow=True)

plt.title('Missing Values',
          color='Green',
          loc='center',
        );
```



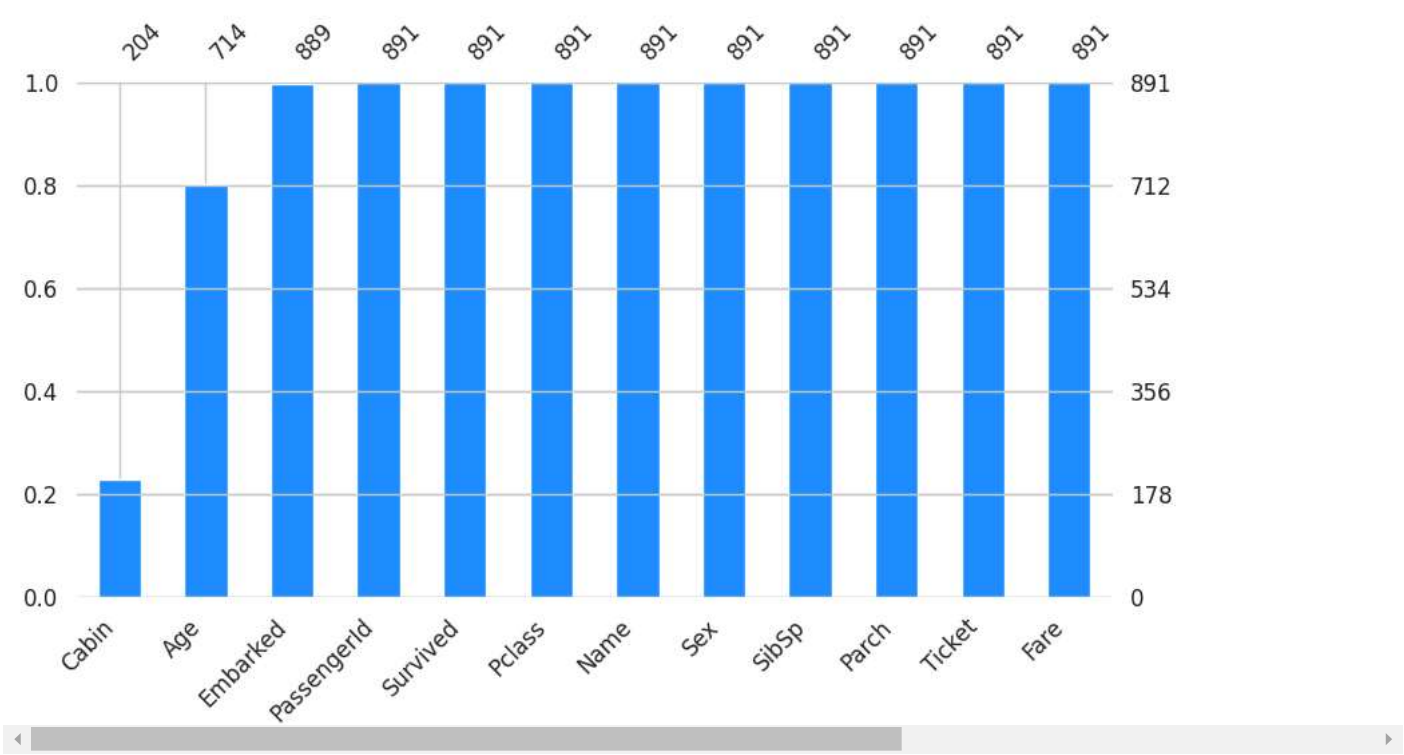

```
sns.heatmap(data.isnull(),cmap='cool');
```



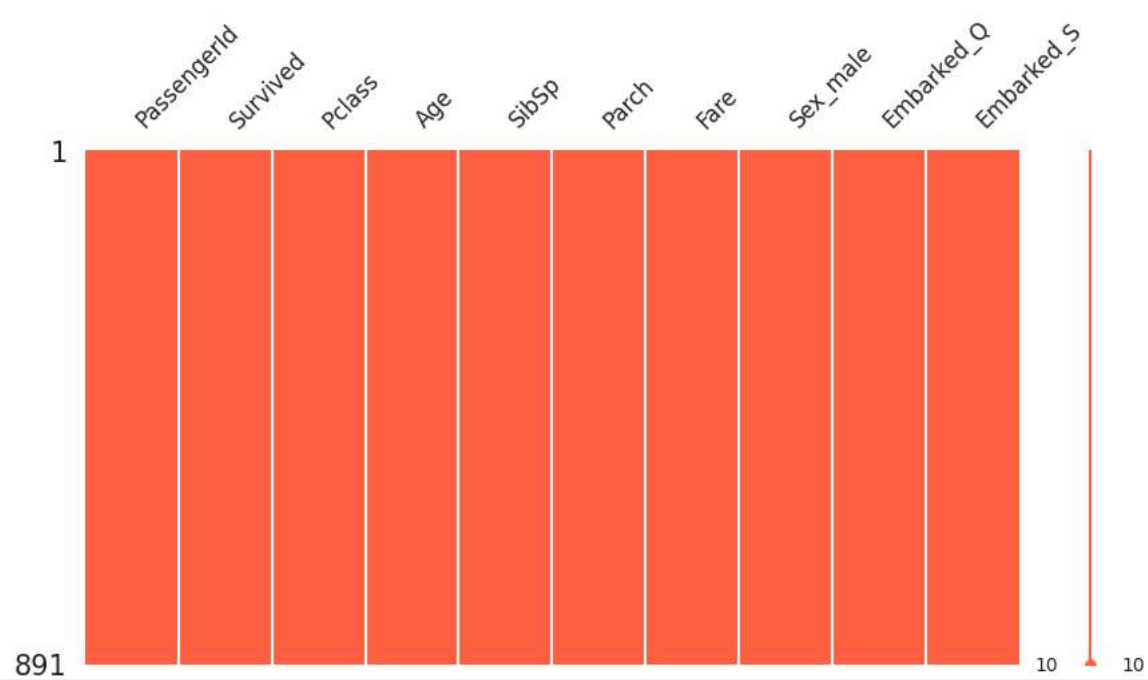

```
import missingno as msno
msno.matrix(data)
plt.show()
```




```
import missingno
missingno.bar(data, color="dodgerblue", sort="ascending", figsize=(10,5), fontsize=12);
```



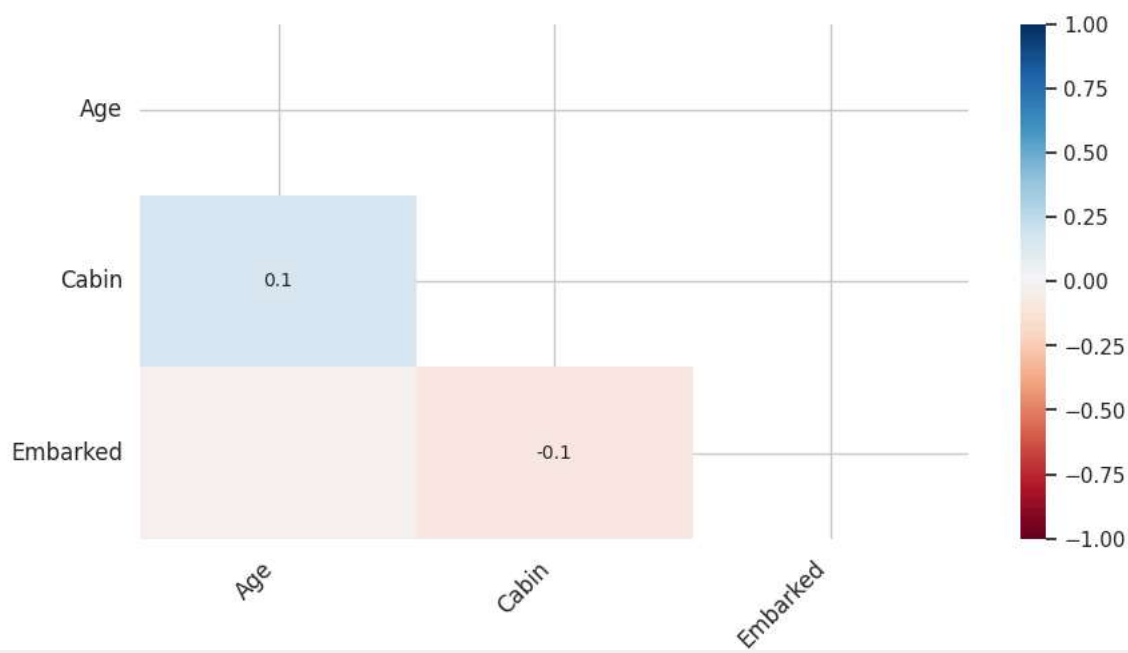

```
missingno.matrix(data, figsize=(10,5), fontsize=12, color=(1, 0.38, 0.27));
```




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

	0
PassengerId	0
Survived	0
Pclass	0
Age	0
SibSp	0
Parch	0
Fare	0
Sex_male	0
Embarked_Q	0
Embarked_S	0

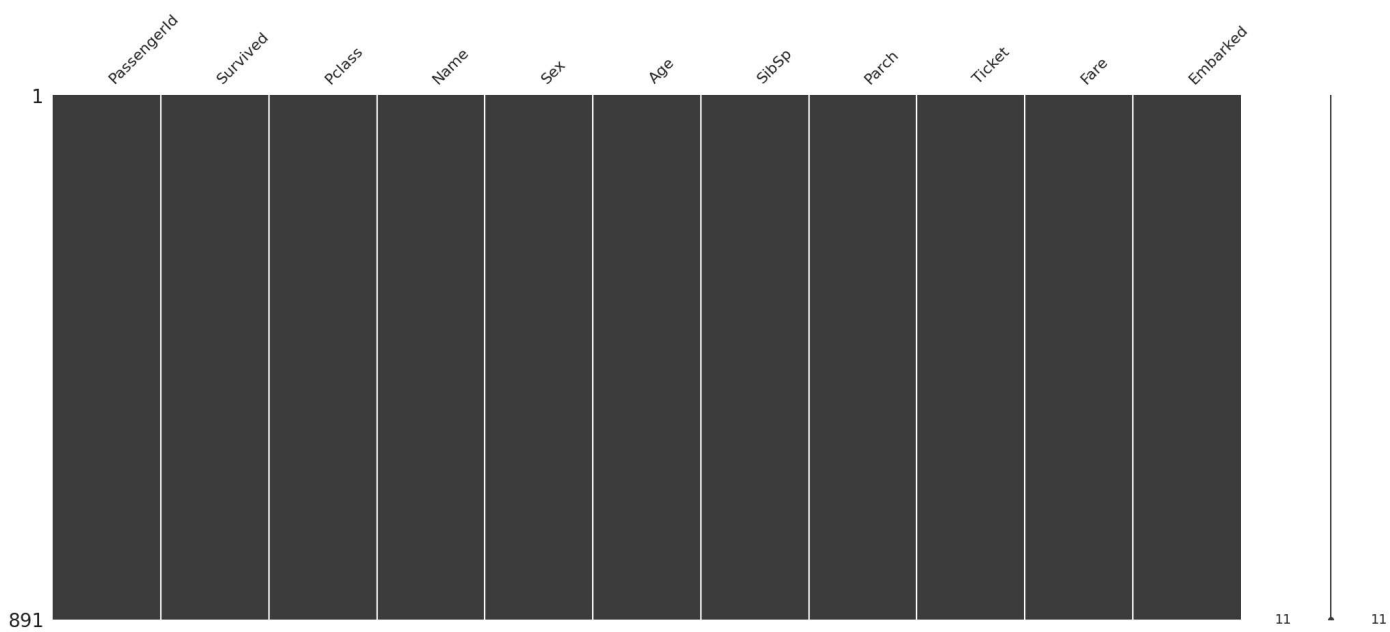
```
missingno.heatmap(data, figsize=(10,5), fontsize=12);
```




```
data['Age'] = data['Age'].fillna(data['Age'].mean())
data[data['Embarked'].isnull()]
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
61	62	1	1	Icard, Miss. Amelie	female	38.0	0	0	113572	80.0	B28	NaN
829	830	1	1	Stone, Mrs. George Nelson (Martha Evelyn)	female	62.0	0	0	113572	80.0	B28	NaN

```
data['Embarked'] = data['Embarked'].fillna(method='bfill')
data = data.drop(['Cabin'],axis=1)
import missingno as msno
msno.matrix(data)
plt.show()
```




```
data = data.drop(['Name','Ticket'],axis=1)
data.head()
```

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	1	0	3	male	22.0	1	0	7.2500	S
1	2	1	1	female	38.0	1	0	71.2833	C
2	3	1	3	female	26.0	0	0	7.9250	S
3	4	1	1	female	35.0	1	0	53.1000	S
4	5	0	3	male	35.0	0	0	8.0500	S

```
data = pd.get_dummies(data,columns=['Sex','Embarked'],drop_first=True)
data.head()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	Sex_male	Embarked_Q	Embarked_S
0	1	0	3	22.0	1	0	7.2500	True	False	True
1	2	1	1	38.0	1	0	71.2833	False	False	False
2	3	1	3	26.0	0	0	7.9250	False	False	True
3	4	1	1	35.0	1	0	53.1000	False	False	True
4	5	0	3	35.0	0	0	8.0500	True	False	True

```
X = data.drop(['Survived'],axis=1)
y = data['Survived']
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2,random_state=21)


from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

X_train = pd.DataFrame(X_train, columns=X.columns
```

```
X_test = pd.DataFrame(X_test, columns=X.columns)
display(X_train.head())
```

	PassengerId	Pclass	Age	SibS	Parch	Fare	Sex_male	Embarked_Q	Embarked_S
0	1.360492	-1.584396	0.010681	-0.47969	-0.460682	-0.018600	0.728823	-0.311564	-1.611198
1	-1.632266	-1.584396	-0.119643	-0.47969	-0.460682	0.079245	0.728823	-0.311564	0.620656
2	-1.344650	-1.584396	-0.503148	-0.479698	0.810657	0.646624	0.728823	-0.311564	-1.611198
3	-1.686680	-0.381742	-1.193456	0.493365	-0.460682	-0.031329	-1.372075	-0.311564	-1.611198
4	-1.111449	0.820913	0.033758	-0.479698	-0.460682	-0.479818	0.728823	-0.311564	0.620656

	PassengerId	Pclass	Age	SibSp	Parch	Fare	Sex_male	Embarked_Q	Embarked_S
0	0.676433	0.820913	-0.273045	0.493365	-0.460682	-0.315867	-1.372075	-0.311564	0.620656
1	-0.248601	0.820913	-0.809952	-0.479698	-0.460682	-0.485419	0.728823	-0.311564	0.620656
2	1.096196	0.820913	-0.733251	-0.479698	-0.460682	-0.467343	0.728823	-0.311564	0.620656
3	1.488753	0.820913	0.010681	-0.479698	-0.460682	0.506858	0.728823	-0.311564	0.620656
4	0.027354	-0.381742	0.493964	0.493365	2.081997	-0.078596	0.728823	-0.311564	0.620656

```
print(X_train.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 712 entries, 0 to 711
Data columns (total 9 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  712 non-null    float64
 1   Pclass       712 non-null    float64
 2   Age          712 non-null    float64
 3   SibSp        712 non-null    float64
 4   Parch        712 non-null    float64
 5   Fare         712 non-null    float64
 6   Sex_male     712 non-null    float64
 7   Embarked_Q   712 non-null    float64
 8   Embarked_S   712 non-null    float64
dtypes: float64(9)
memory usage: 50.2 KB
None
```

```
print(X_test.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 179 entries, 0 to 178
Data columns (total 9 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  179 non-null    float64
 1   Pclass       179 non-null    float64
 2   Age          179 non-null    float64
 3   SibSp        179 non-null    float64
 4   Parch        179 non-null    float64
 5   Fare         179 non-null    float64
 6   Sex_male     179 non-null    float64
 7   Embarked_Q   179 non-null    float64
 8   Embarked_S   179 non-null    float64
dtypes: float64(9)
memory usage: 12.7 KB
None
```