```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

```
df=pd.read_csv('train.csv')
```

```
df.head()
```

	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
				Futrelle, Mrs. Jacques Heath (Lily								

Univariate Analysis

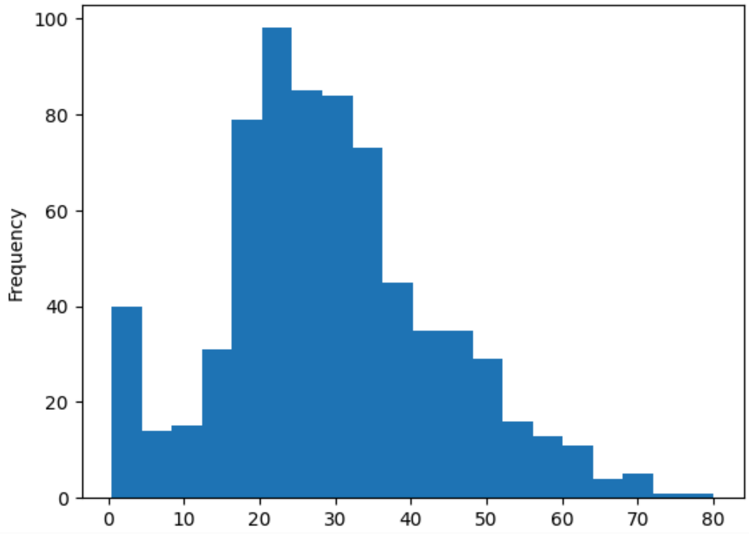
```
df['Age'].describe()
```

	Age
count	714.000000
mean	29.699118
std	14.526497
min	0.420000
25%	20.125000
50%	28.000000
75%	38.000000
max	80.000000

dtype: float64

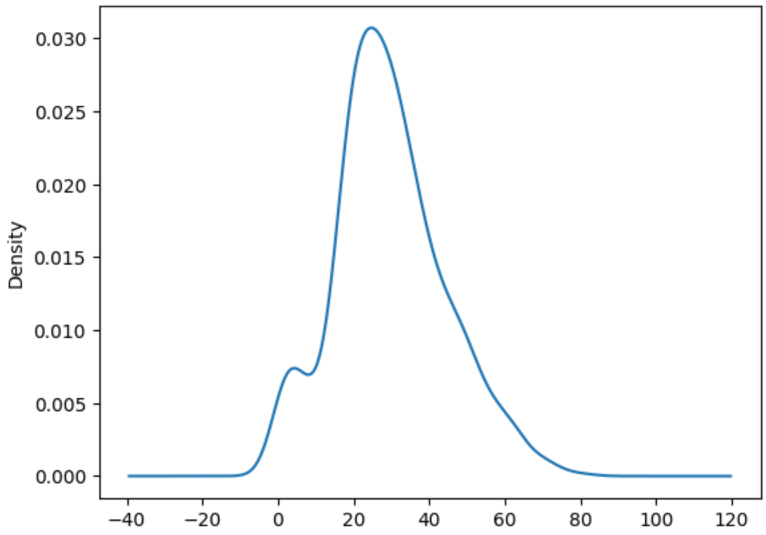
```
df['Age'].plot(kind='hist',bins=20)
```

```
<Axes: ylabel='Frequency'>
```



```
df['Age'].plot(kind='kde')
```

```
<Axes: ylabel='Density'>
```



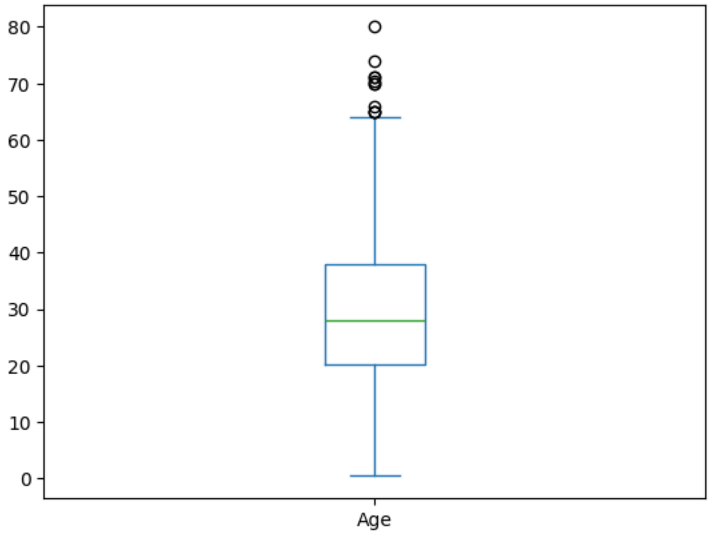
```
print(df['Age'].skew())
```

```
0.38910778230082704
```

Start coding or generate with AI.

```
df['Age'].plot(kind='box')
```

```
<Axes: >
```



```
df[df['Age'] > 65]
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
33	34	0	2	Wheadon, Mr. Edward H	male	66.0	0	0	C.A. 24579	10.5000	NaN	S
96	97	0	1	Goldschmidt, Mr. George B	male	71.0	0	0	PC 17754	34.6542	A5	C
116	117	0	3	Connors, Mr. Patrick	male	70.5	0	0	370369	7.7500	NaN	Q
493	494	0	1	Artagaveytia, Mr. Ramon	male	71.0	0	0	PC 17609	49.5042	NaN	C
630	631	1	1	Barkworth, Mr. Algernon Henry Wilson	male	80.0	0	0	27042	30.0000	A23	S
672	673	0	2	Mitchell, Mr. Henry Michael	male	70.0	0	0	C.A. 24580	10.5000	NaN	S
745	746	0	1	Crosby, Capt. Edward Gifford	male	70.0	1	1	WE/P 5735	71.0000	B22	S
851	852	0	3	Svensson, Mr. Johan	male	74.0	0	0	347060	7.7750	NaN	S

```
print(df['Age'].isnull().sum()/len(df['Age']))
```

```
0.19865319865319866
```

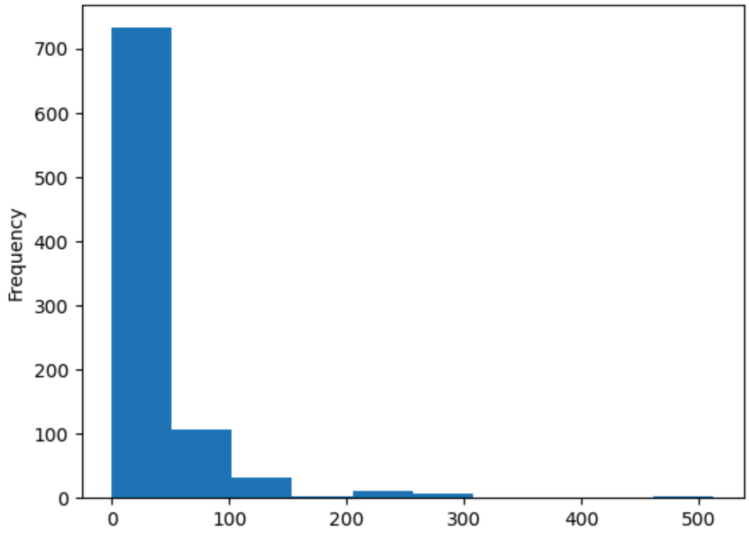
```
df['Fare'].describe()
```

```
             Fare
count  891.000000
 mean   32.204208
  std   49.693429
  min    0.000000
  25%    7.910400
  50%   14.454200
  75%   31.000000
  max  512.329200

dtype: float64
```

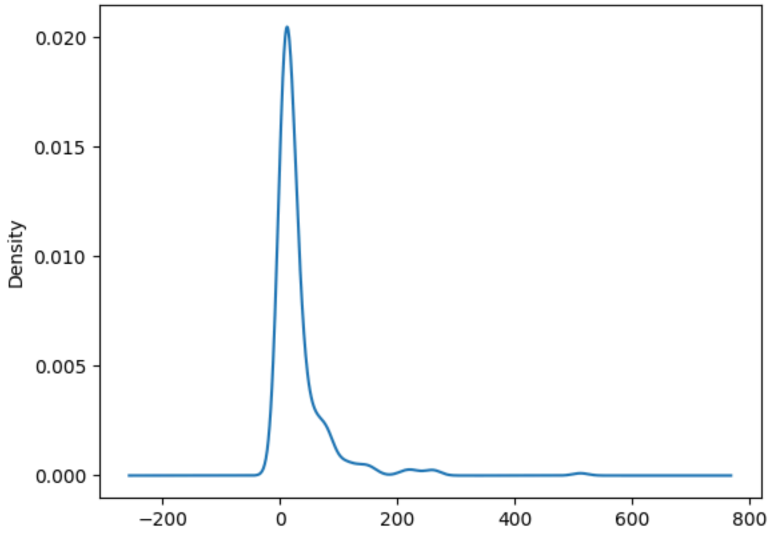
```
df['Fare'].plot(kind='hist')
```

```
<Axes: ylabel='Frequency'>
```



```
df['Fare'].plot(kind='kde')
```

```
<Axes: ylabel='Density'>
```

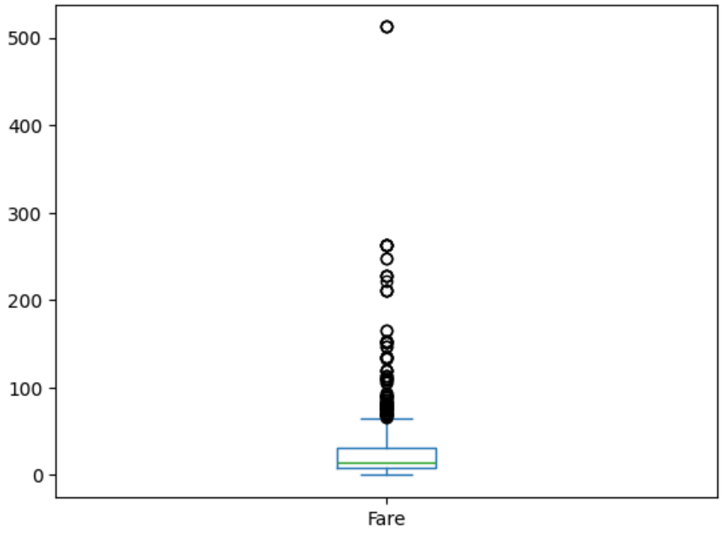


```
print(df['Fare'].skew())
```

```
4.787316519674893
```

```
df['Fare'].plot(kind='box')
```

```
<Axes: >
```



```
df[df['Fare'] > 250]
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
27	28	0	1	Fortune, Mr. Charles Alexander	male	19.0	3	2	19950	263.0000	C23 C25 C27	S
88	89	1	1	Fortune, Miss. Mabel Helen	female	23.0	3	2	19950	263.0000	C23 C25 C27	S
258	259	1	1	Ward, Miss. Anna	female	35.0	0	0	PC 17755	512.3292	NaN	C
311	312	1	1	Ryerson, Miss. Emily Borie	female	18.0	2	2	PC 17608	262.3750	B57 B59 B63 B66	C
341	342	1	1	Fortune, Miss. Alice Elizabeth	female	24.0	3	2	19950	263.0000	C23 C25 C27	S
438	439	0	1	Fortune, Mr. Mark	male	64.0	1	4	19950	263.0000	C23 C25 C27	S
679	680	1	1	Cardeza, Mr. Thomas Drake Martinez	male	36.0	0	1	PC 17755	512.3292	B51 B53 B55	C

```
print(df['Fare'].isnull().sum())
```

```
0
```

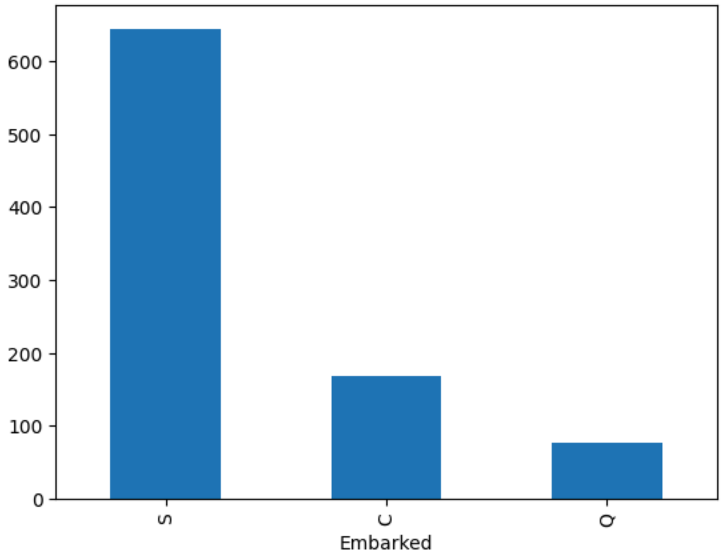
```
df['Embarked'].value_counts()
```

	count
Embarked	
S	644
C	168
Q	77

dtype: int64

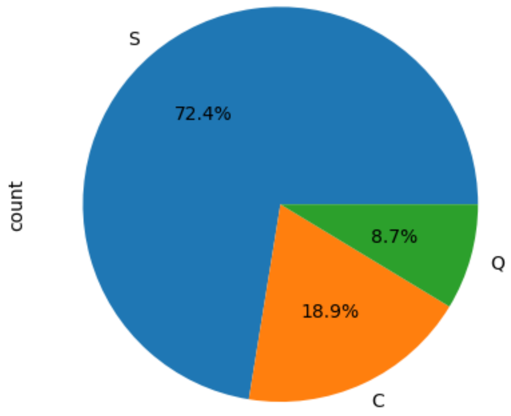
```
df['Embarked'].value_counts().plot(kind='bar')
```

```
<Axes: xlabel='Embarked'>
```



```
df['Embarked'].value_counts().plot(kind='pie',autopct='%0.1f%%')
```

```
<Axes: ylabel='count'>
```



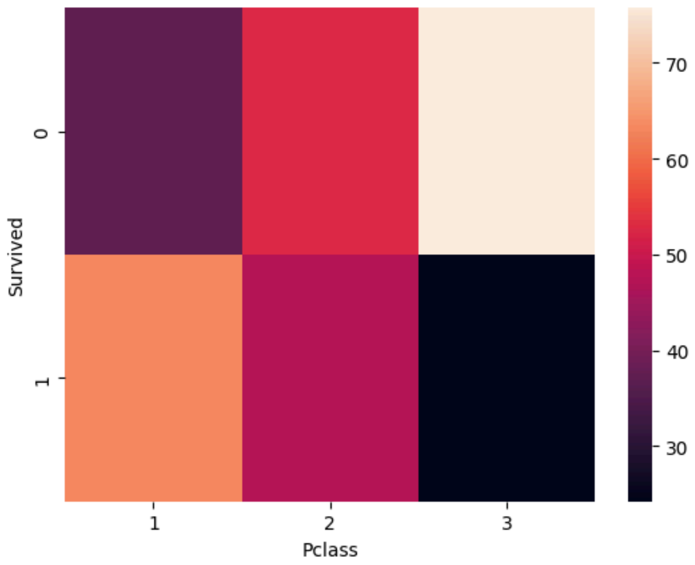
```
print(df['Sex'].isnull().sum())
```

```
0
```

Bivariate Analysis

```
sns.heatmap(pd.crosstab(df['Survived'],df['Pclass'],normalize='columns')*100)
```

```
<Axes: xlabel='Pclass', ylabel='Survived'>
```



```
pd.crosstab(df['Survived'],df['Sex'],normalize='columns')*100
```

Sex	female	male
Survived		
0	25.796178	81.109185
1	74.203822	18.890815

```
pd.crosstab(df['Survived'],df['Embarked'],normalize='columns')*100
```

Embarked	C	Q	S
Survived			
0	44.642857	61.038961	66.304348
1	55.357143	38.961039	33.695652

```
pd.crosstab(df['Sex'],df['Embarked'],normalize='columns')*100
```

Embarked	C	Q	S
Sex			
female	43.452381	46.753247	31.521739
male	56.547619	53.246753	68.478261

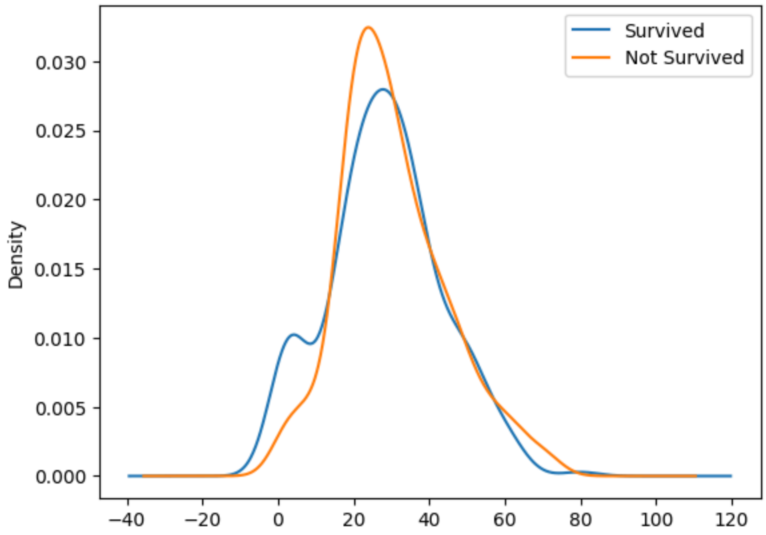
```
pd.crosstab(df['Pclass'],df['Embarked'],normalize='columns')*100
```

Embarked	C	Q	S
Pclass			
1	50.595238	2.597403	19.720497
2	10.119048	3.896104	25.465839
3	39.285714	93.506494	54.813665

```
# survived and age

df[df['Survived'] == 1]['Age'].plot(kind='kde',label='Survived')
df[df['Survived'] == 0]['Age'].plot(kind='kde',label='Not Survived')

plt.legend()
plt.show()
```



```
print(df[df['Pclass'] == 1]['Age'].mean())
```

```
38.233440860215055
```

```
# Feature Engineering on Fare col
df['SibSp'].value_counts()
```

SibSp	count
0	608
1	209
2	28
4	18
3	16
8	7
5	5

dtype: int64

```
df[df['Ticket'] == 'CA. 2343']
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
159	160	0	3	Sage, Master. Thomas Henry	male	NaN	8	2	CA. 2343	69.55	NaN	S
180	181	0	3	Sage, Miss. Constance Gladys	female	NaN	8	2	CA. 2343	69.55	NaN	S
201	202	0	3	Sage, Mr. Frederick	male	NaN	8	2	CA. 2343	69.55	NaN	S
324	325	0	3	Sage, Mr. George John Jr	male	NaN	8	2	CA. 2343	69.55	NaN	S
792	793	0	3	Sage, Miss. Stella Anna	female	NaN	8	2	CA. 2343	69.55	NaN	S
846	847	0	3	Sage, Mr. Douglas Bullen	male	NaN	8	2	CA. 2343	69.55	NaN	S
863	864	0	3	Sage, Miss. Dorothy Edith "Dolly"	female	NaN	8	2	CA. 2343	69.55	NaN	S

```
df[df['Name'].str.contains('Sage')]
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
159	160	0	3	Sage, Master. Thomas Henry	male	NaN	8	2	CA. 2343	69.55	NaN	S
180	181	0	3	Sage, Miss. Constance Gladys	female	NaN	8	2	CA. 2343	69.55	NaN	S
201	202	0	3	Sage, Mr. Frederick	male	NaN	8	2	CA. 2343	69.55	NaN	S
324	325	0	3	Sage, Mr. George John Jr	male	NaN	8	2	CA. 2343	69.55	NaN	S
641	642	1	1	Sagesser, Mlle. Emma	female	24.0	0	0	PC 17477	69.30	B35	C
792	793	0	3	Sage, Miss. Stella Anna	female	NaN	8	2	CA. 2343	69.55	NaN	S
846	847	0	3	Sage, Mr. Douglas Bullen	male	NaN	8	2	CA. 2343	69.55	NaN	S
863	864	0	3	Sage, Miss. Dorothy Edith "Dolly"	female	NaN	8	2	CA. 2343	69.55	NaN	S

```
df1 = pd.read_csv('test.csv')
```

```
df = pd.concat([df,df1])
```

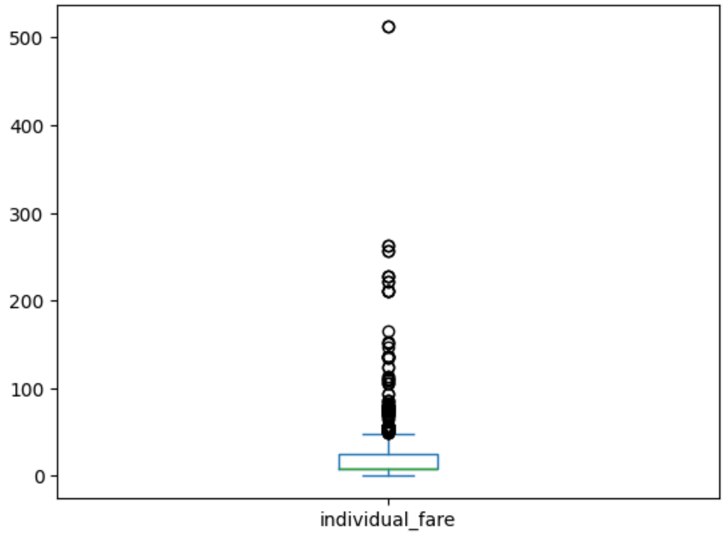
```
df[df['Ticket'] == 'CA 2144']
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
59	60	0.0	3	Goodwin, Master. William Frederick	male	11.0	5	2	CA 2144	46.9	NaN	S
71	72	0.0	3	Goodwin, Miss. Lillian Amy	female	16.0	5	2	CA 2144	46.9	NaN	S
386	387	0.0	3	Goodwin, Master. Sidney Leonard	male	1.0	5	2	CA 2144	46.9	NaN	S
480	481	0.0	3	Goodwin, Master. Harold Victor	male	9.0	5	2	CA 2144	46.9	NaN	S
678	679	0.0	3	Goodwin, Mrs. Frederick (Augusta Tyler)	female	43.0	1	6	CA 2144	46.9	NaN	S
683	684	0.0	3	Goodwin, Mr. Charles Edward	male	14.0	5	2	CA 2144	46.9	NaN	S
139	1031	NaN	3	Goodwin, Mr. Charles Frederick	male	40.0	1	6	CA 2144	46.9	NaN	S
140	1032	NaN	3	Goodwin, Miss. Jessie Allis	female	10.0	5	2	CA 2144	46.9	NaN	S

```
df['individual_fare'] = df['Fare']/(df['SibSp'] + df['Parch'] + 1)
```

```
df['individual_fare'].plot(kind='box')
```

```
<Axes: >
```



```
df[['individual_fare','Fare']].describe()
```

	individual_fare	Fare
count	1308.000000	1308.000000
mean	20.518215	33.295479
std	35.774337	51.758668
min	0.000000	0.000000
25%	7.452767	7.895800
50%	8.512483	14.454200
75%	24.237500	31.275000
max	512.329200	512.329200

```
df['Fare']
```

	Fare
0	7.2500
1	71.2833
2	7.9250
3	53.1000
4	8.0500
...	...
413	8.0500
414	108.9000
415	7.2500
416	8.0500
417	22.3583

1309 rows × 1 columns

dtype: float64

```
df
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	individual_fare
0	1	0.0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	3.625000
1	2	1.0	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	35.641650
2	3	1.0	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	7.925000
3	4	1.0	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	26.550000
4	5	0.0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	8.050000
...	...	...	...	...	...	...	...	...	...	...	...	...	...

```
df['family_size'] = df['SibSp'] + df['Parch'] + 1
```

```
# family_type
# 1 -> alone
# 2-4 -> small
# >5 -> large

def transform_family_size(num):

  if num == 1:
    return 'alone'
  elif num>1 and num <5:
    return "small"
  else:
    return "large"
```

```
df['family_type'] = df['family_size'].apply(transform_family_size)
```

```
df
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	individual_fare	fami
0	1	0.0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	3.625000	
1	2	1.0	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	35.641650	
2	3	1.0	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	7.925000	
3	4	1.0	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	26.550000	
4	5	0.0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	8.050000	

```
pd.crosstab(df['Survived'],df['family_type'],normalize='columns')*100
```

family_type	alone	large	small
Survived			
0.0	69.646182	83.870968	42.123288
1.0	30.353818	16.129032	57.876712

```
df['surname'] = df['Name'].str.split(',').str.get(0)
```

```
df
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	individual_fare	fami
0	1	0.0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	3.625000	
1	2	1.0	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	35.641650	
2	3	1.0	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	7.925000	
3	4	1.0	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	26.550000	
4	5	0.0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	8.050000	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	

```
df['title'] = df['Name'].str.split(',').str.get(1).str.strip().str.split(' ').str.get(0)
```

```
temp_df = df[df['title'].isin(['Mr.','Miss.','Mrs.','Master.','ootherr'])]
```

```
pd.crosstab(temp_df['Survived'],temp_df['title'],normalize='columns')*100
```

title	Master.	Miss.	Mr.	Mrs.
Survived				
0.0	42.5	30.21978	84.332689	20.8
1.0	57.5	69.78022	15.667311	79.2

```
df['title'] = df['title'].str.replace('Rev.','other')
df['title'] = df['title'].str.replace('Dr.','other')
df['title'] = df['title'].str.replace('Col.','other')
df['title'] = df['title'].str.replace('Major.','other')
df['title'] = df['title'].str.replace('Capt.','other')
df['title'] = df['title'].str.replace('the','other')
df['title'] = df['title'].str.replace('Jonkheer.','other')
# ,'Dr.','Col.','Major.','Don.','Capt.','the','Jonkheer.']
```

```
print(df['Cabin'].isnull().sum()/len(df['Cabin']))
```

```
0.774637127578304
```

```
df['Cabin'].fillna('M',inplace=True)
```

```
df['Cabin'].value_counts()
```

	count
Cabin	
M	1014
C23 C25 C27	6
G6	5
B57 B59 B63 B66	5
F33	4
...	...
C39	1
B24	1
D40	1
D38	1
C105	1

187 rows × 1 columns

dtype: int64

```
df['deck'] = df['Cabin'].str[0]
```

```
df['deck'].value_counts()
```

	count
deck	
M	1014
C	94
B	65
D	46
E	41
A	22
F	21
G	5
T	1

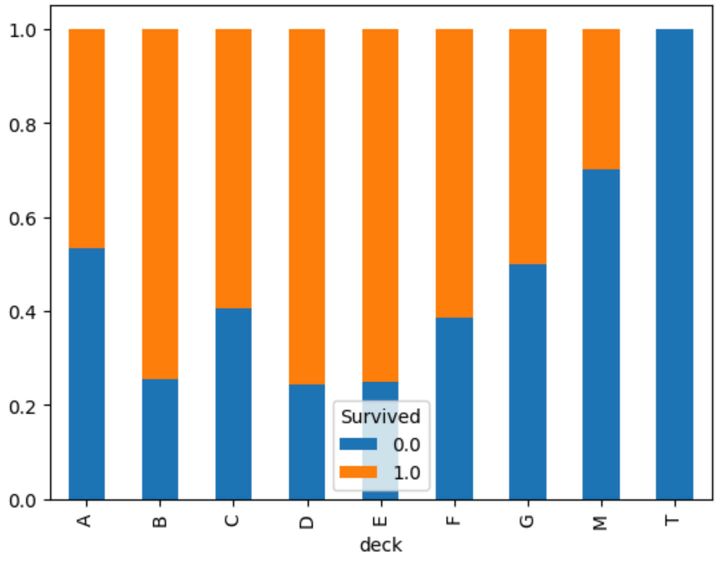
dtype: int64

```
pd.crosstab(df['deck'],df['Pclass'])
```

Pclass	1	2	3
deck			
A	22	0	0
B	65	0	0
C	94	0	0
D	40	6	0
E	34	4	3
F	0	13	8
G	0	0	5
M	67	254	693
T	1	0	0

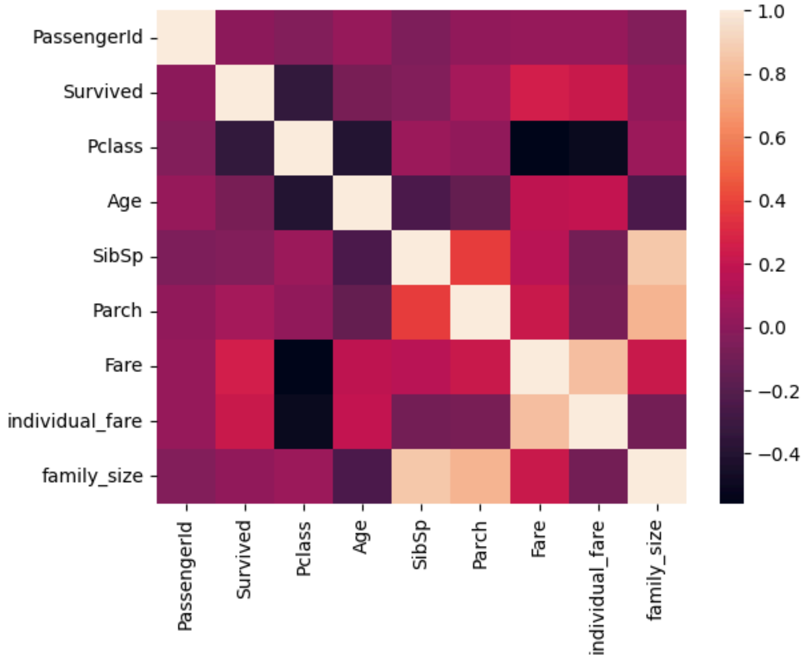
```
pd.crosstab(df['deck'],df['Survived'],normalize='index').plot(kind='bar',stacked=True)
```

```
<Axes: xlabel='deck'>
```



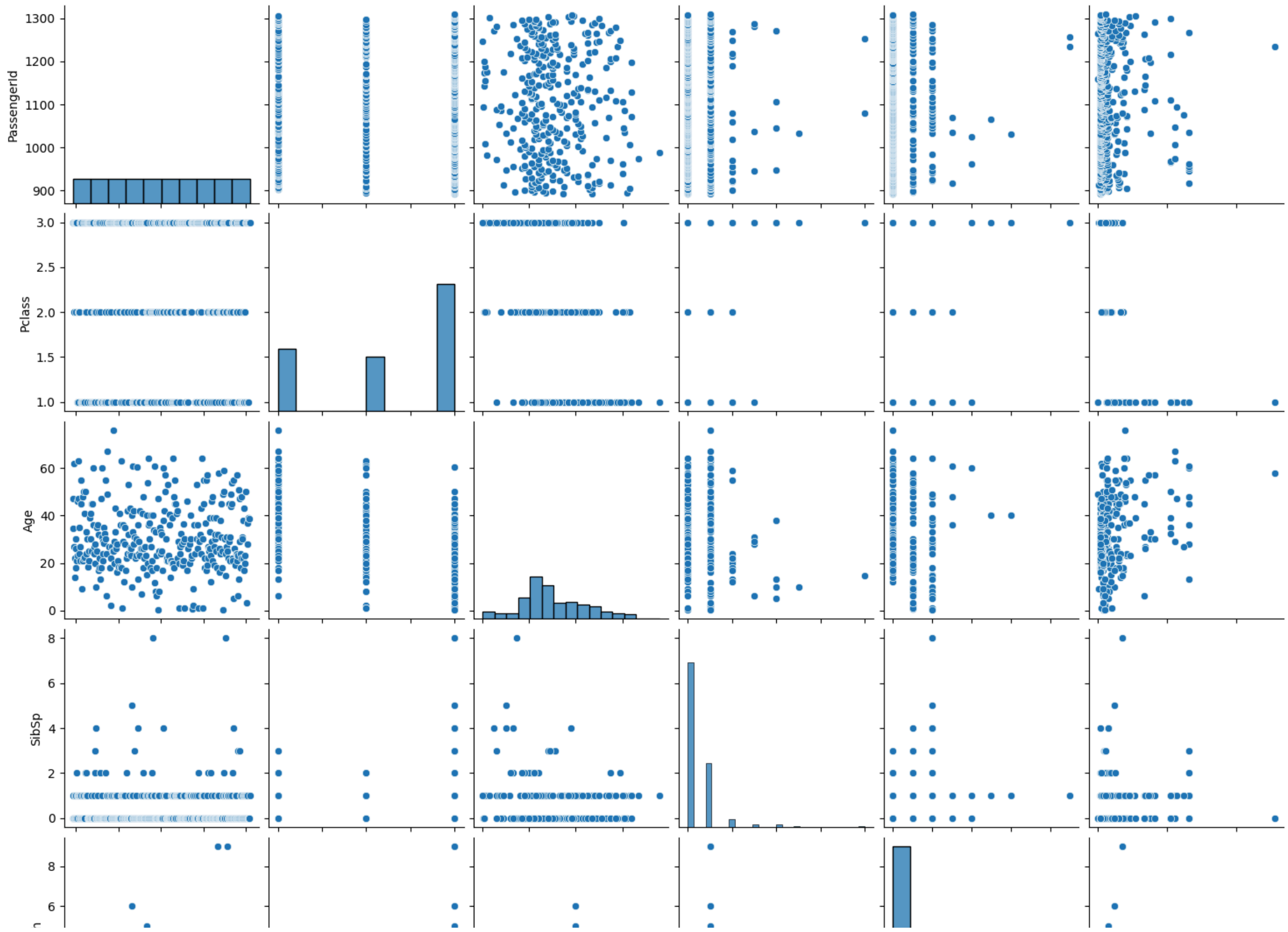
```
sns.heatmap(df.corr(numeric_only=True))
```

```
<Axes: >
```



```
sns.pairplot(df1)
```

```
<seaborn.axisgrid.PairGrid at 0x78fc14689490>
```


PassengerId
Pclass
Age
SibSp