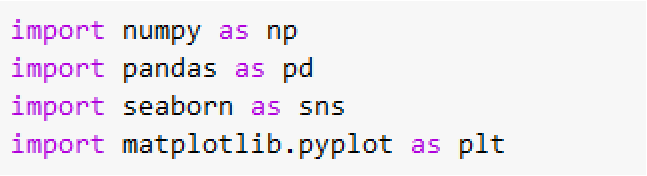
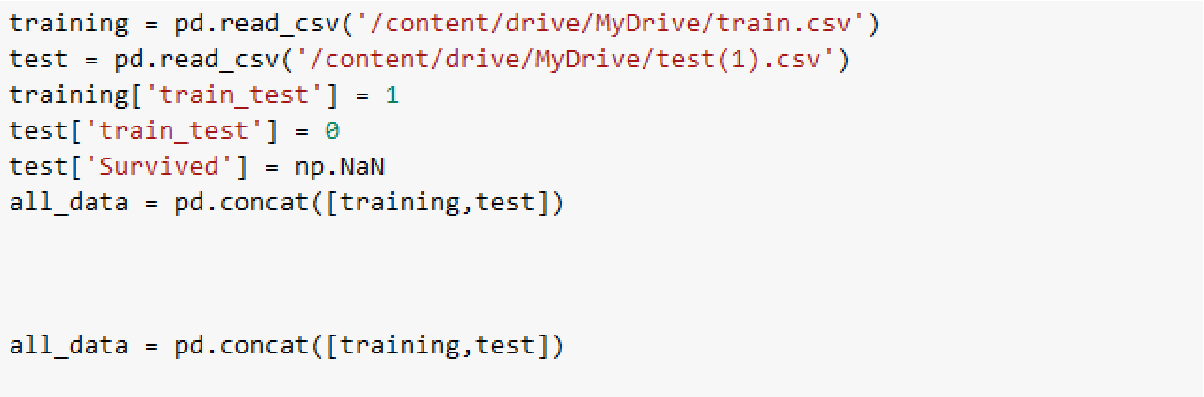
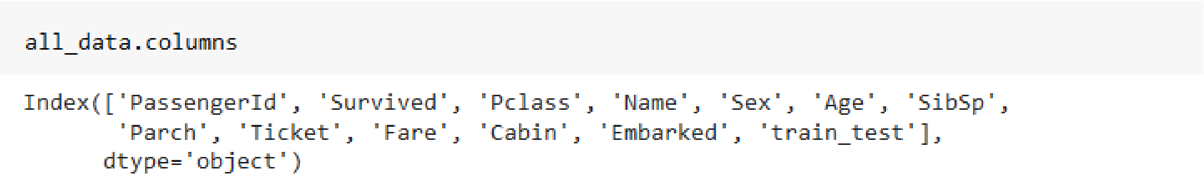


```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

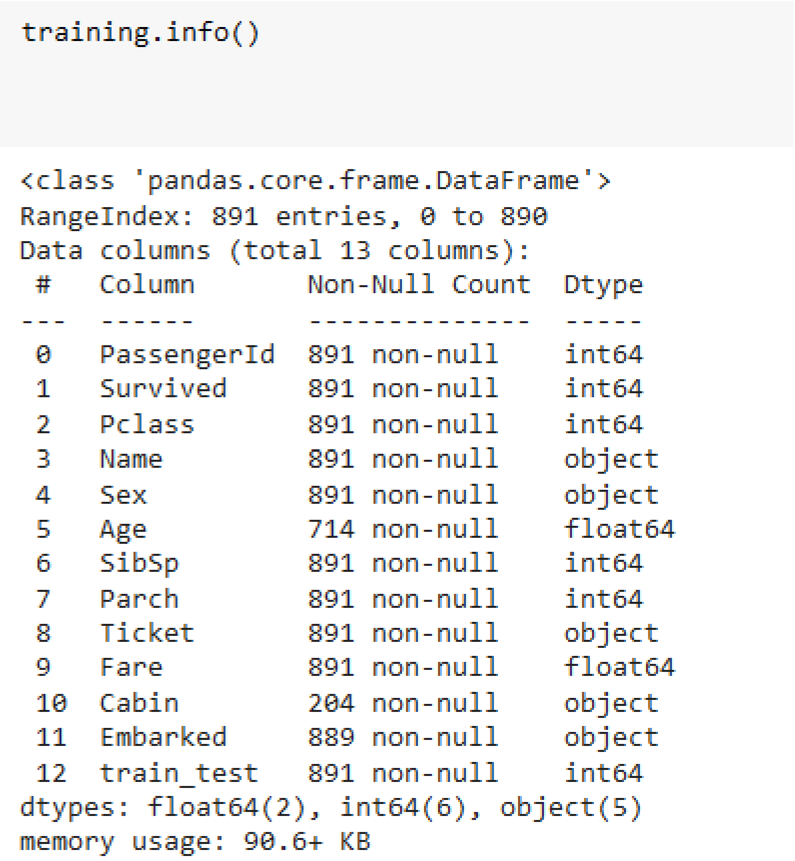


```
training = pd.read_csv('/content/drive/MyDrive/train.csv')
test = pd.read_csv('/content/drive/MyDrive/test(1).csv')
training['train_test'] = 1
test['train_test'] = 0
test['Survived'] = np.NaN
all_data = pd.concat([training,test])



all_data = pd.concat([training,test])
```



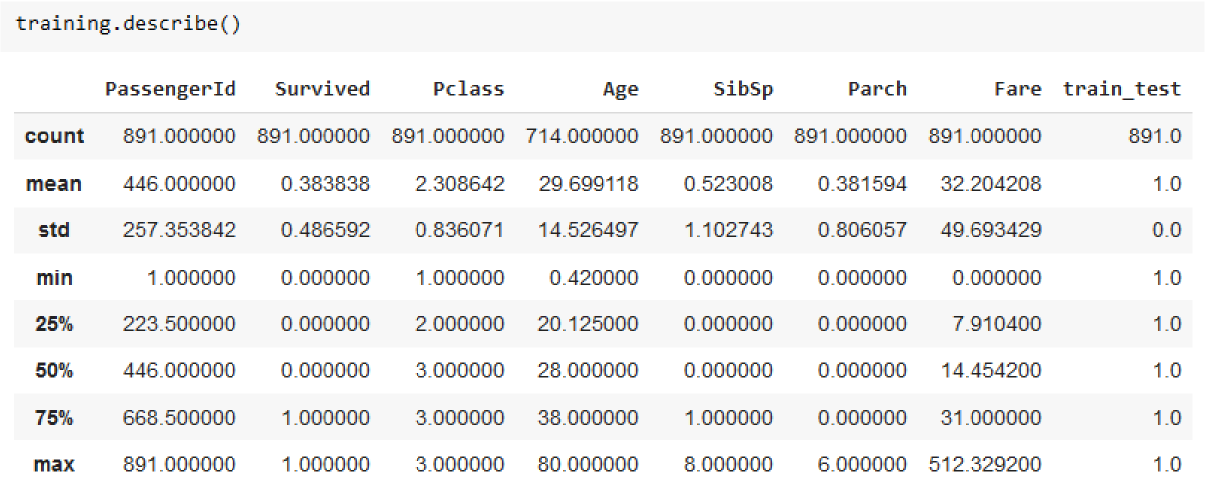
```
all_data.columns
```

```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked', 'train_test'],
      dtype='object')
```



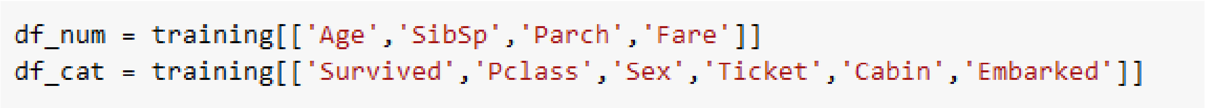
```
training.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 13 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 
 12  train_test   891 non-null    int64  
dtypes: float64(2), int64(6), object(5)
memory usage: 90.6+ KB
```

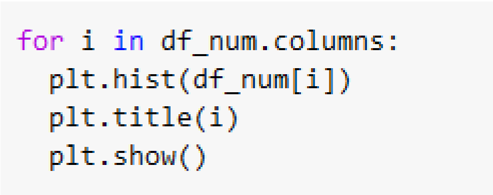


```
training.describe()
```

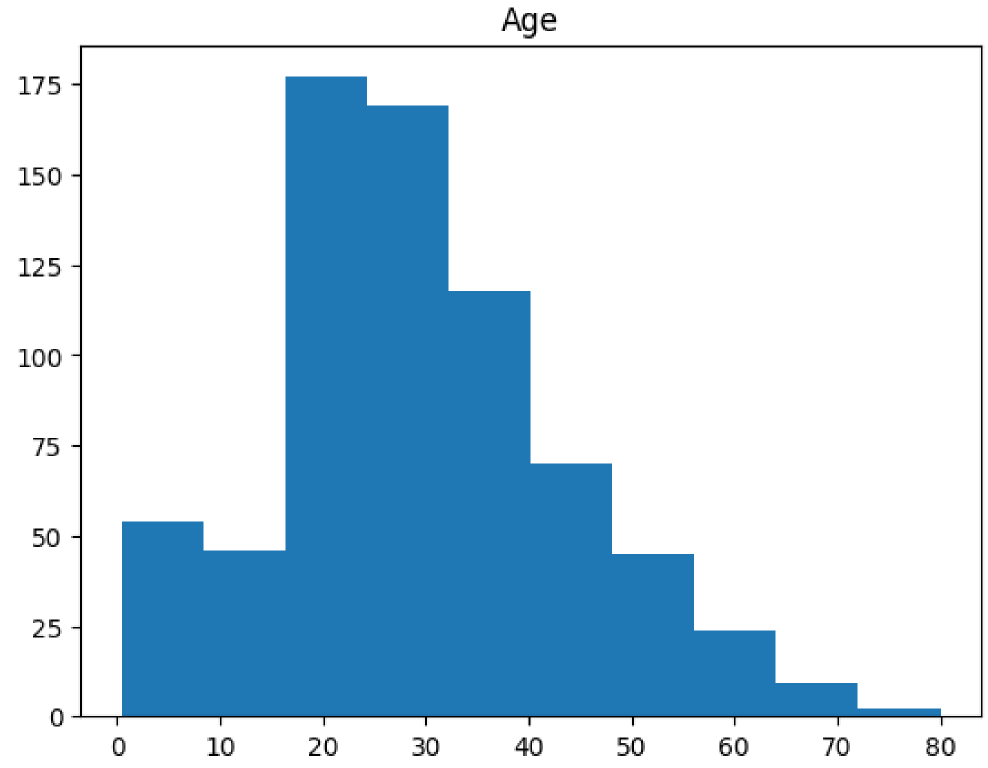
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	train_test
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000	891.0
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208	1.0
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429	0.0
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000	1.0
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400	1.0
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200	1.0
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000	1.0
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200	1.0

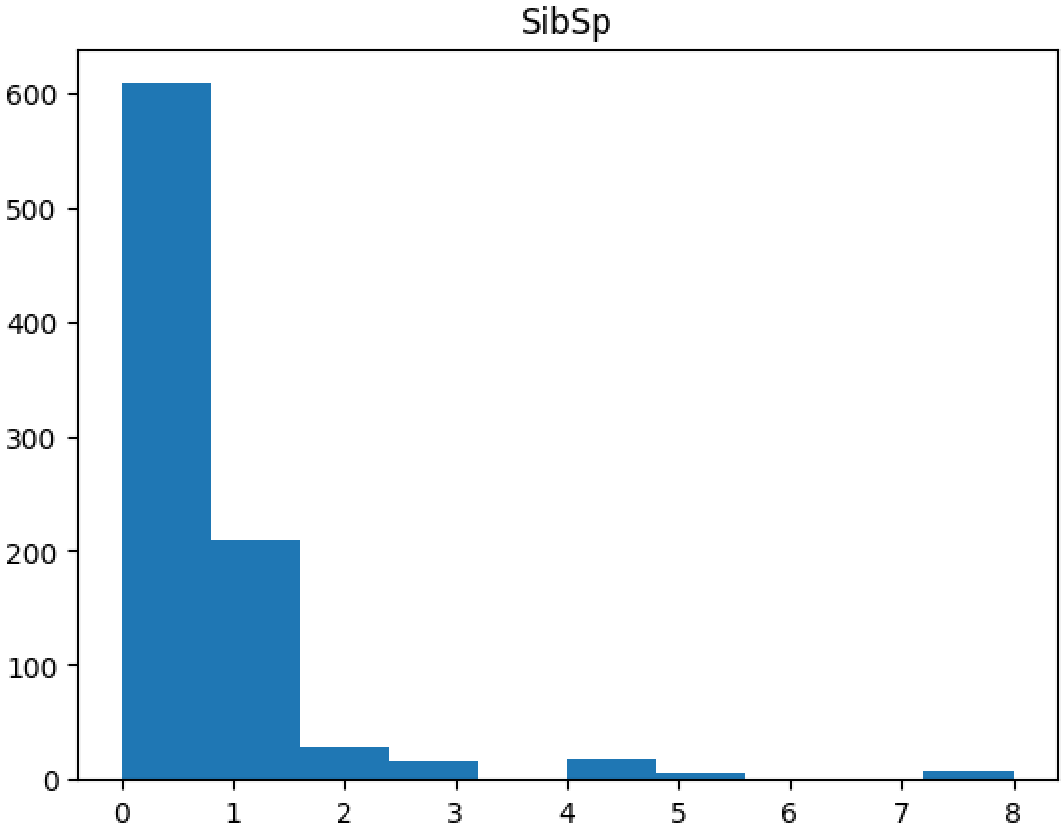


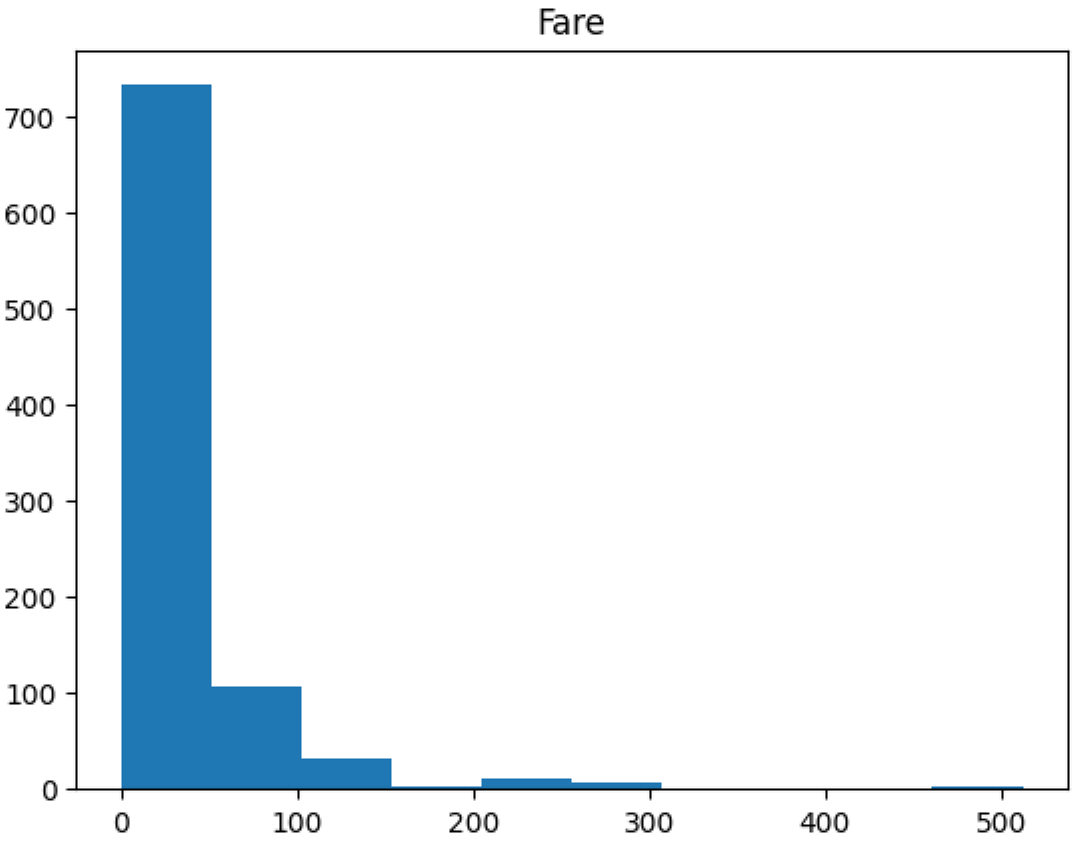
```
df_num = training[['Age','SibSp','Parch','Fare']]
df_cat = training[['Survived','Pclass','Sex','Ticket','Cabin','Embarked']]
```

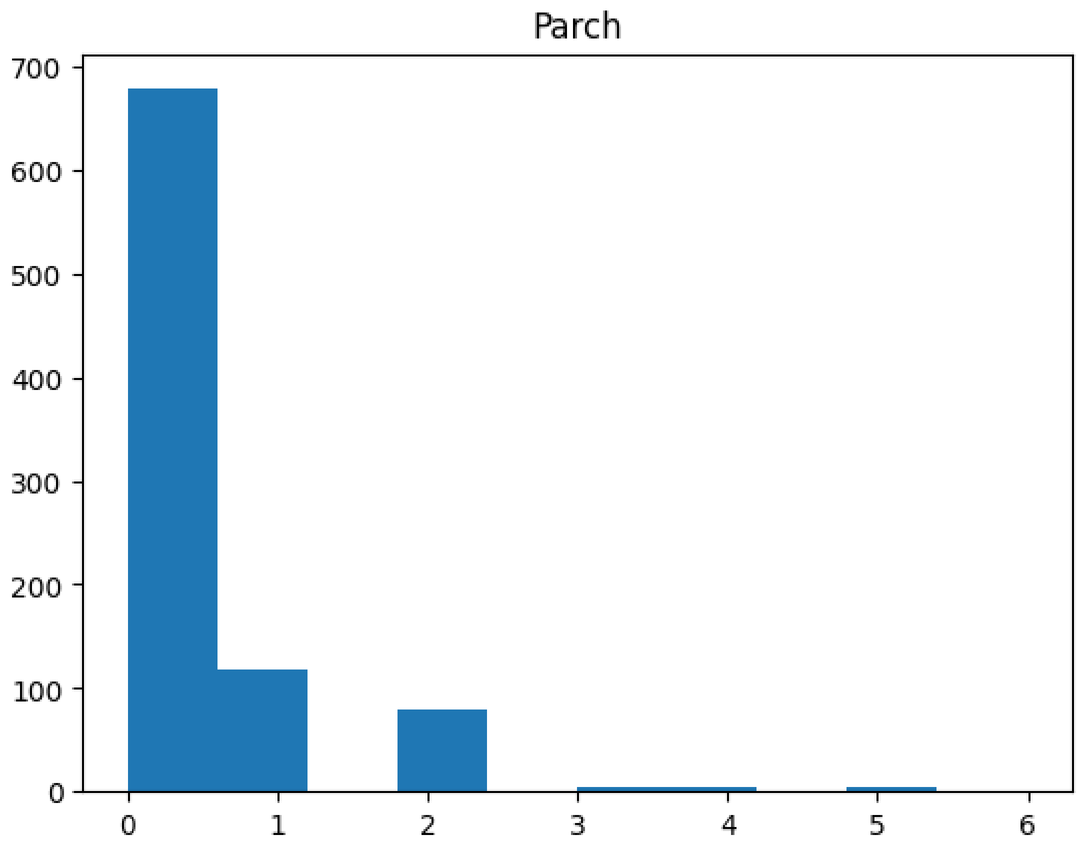


```
for i in df_num.columns:
  plt.hist(df_num[i])
  plt.title(i)
  plt.show()
```



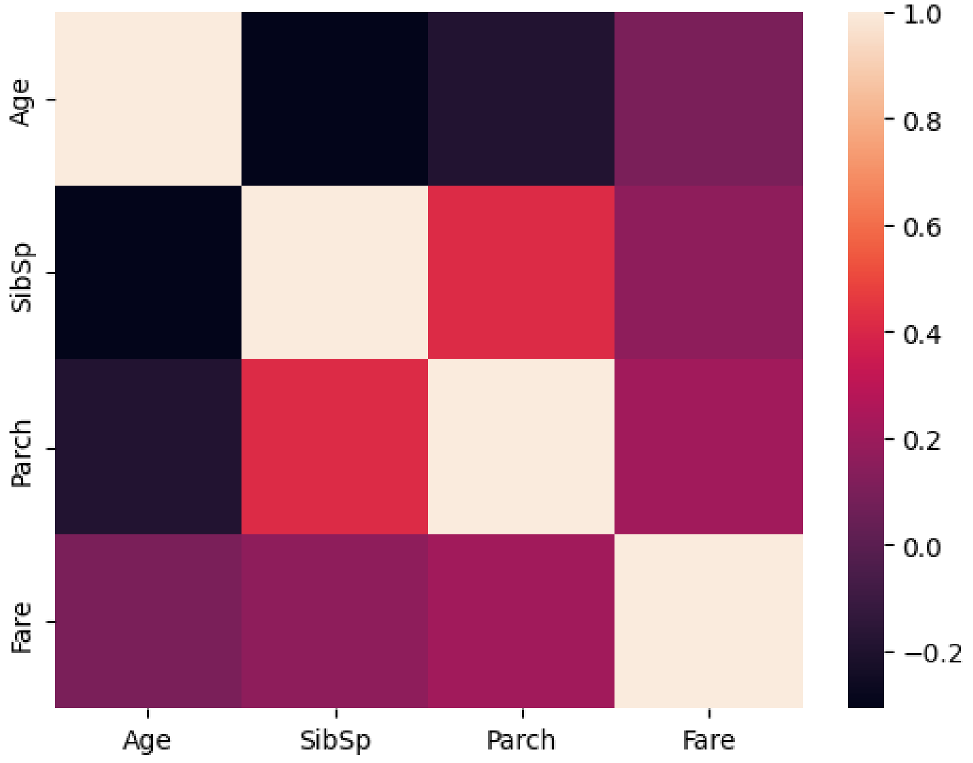

SibSp
600
500
400
300
200
100
0
0
1
2
3
4
5
6
7
8

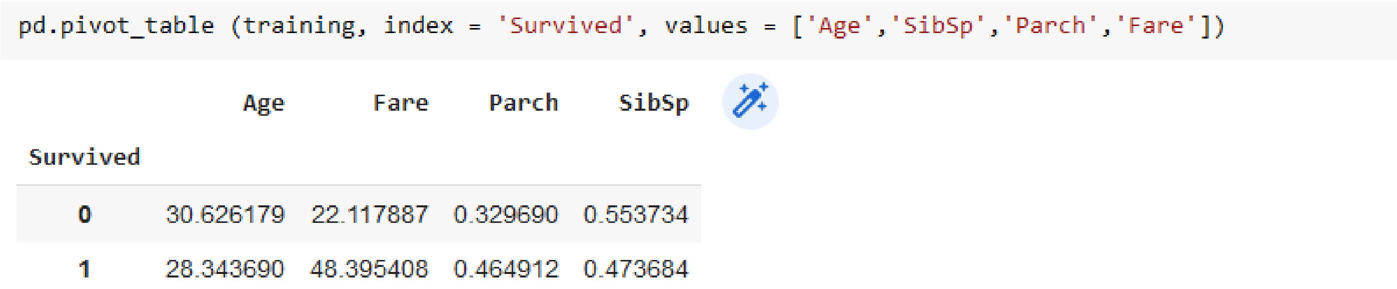

Fare
700
600
500
400
300
200
100
0
0
100
200
300
400
500



```
[ ]  sns.heatmap(df_num.corr())
```

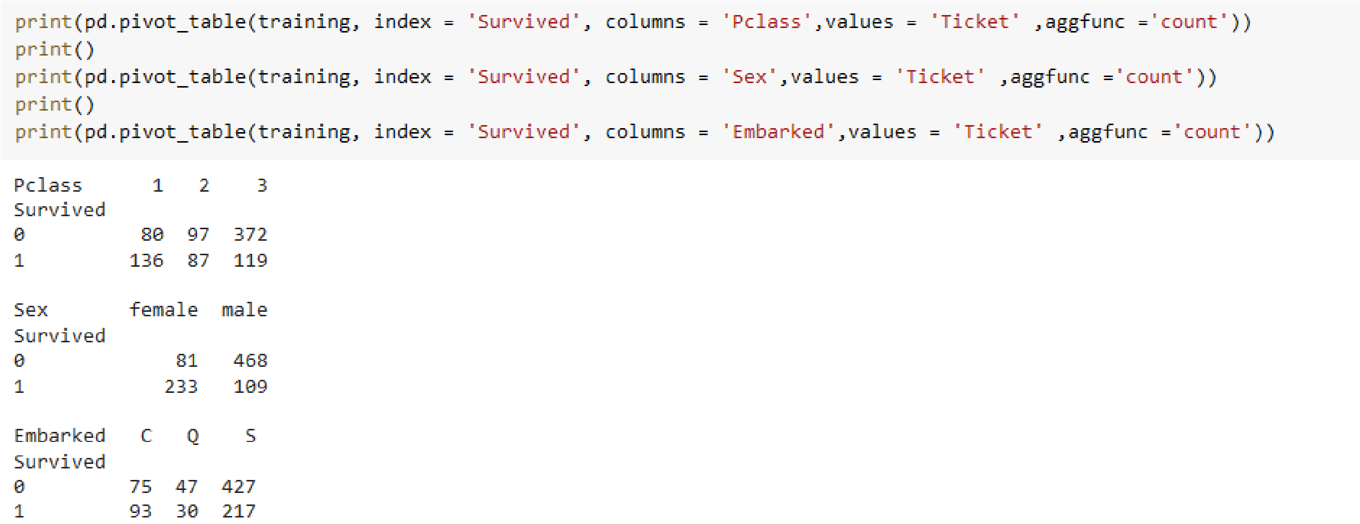
```
<Axes: >
```





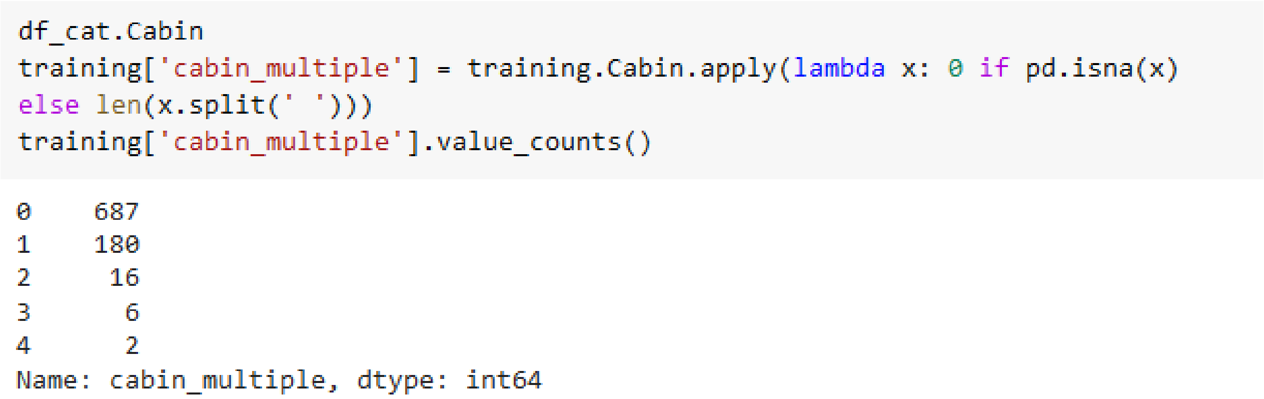
```
pd.pivot_table (training, index = 'Survived', values = ['Age','SibSp','Parch','Fare'])
```

	Age	Fare	Parch	SibSp
Survived				
0	30.626179	22.117887	0.329690	0.553734
1	28.343690	48.395408	0.464912	0.473684



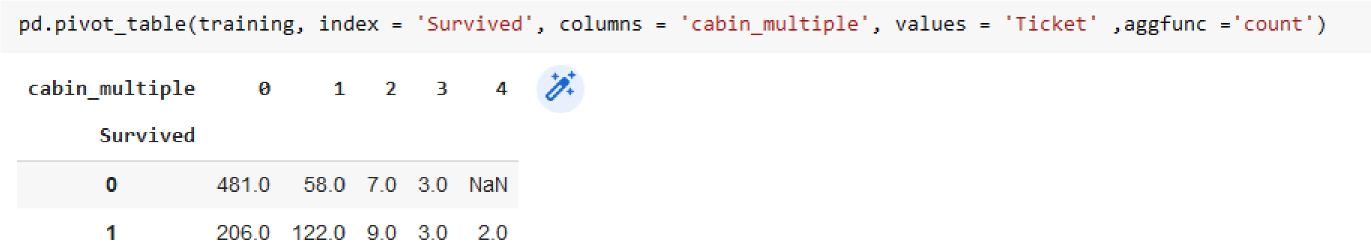
```
print(pd.pivot_table(training, index = 'Survived', columns = 'Pclass',values = 'Ticket' ,aggfunc ='count'))
print()
print(pd.pivot_table(training, index = 'Survived', columns = 'Sex',values = 'Ticket' ,aggfunc ='count'))
print()
print(pd.pivot_table(training, index = 'Survived', columns = 'Embarked',values = 'Ticket' ,aggfunc ='count'))
```

```
Pclass      1   2    3
Survived
0          80  97  372
1         136  87  119

Sex       female  male
Survived
0             81   468
1            233   109

Embarked   C   Q    S
Survived
0         75  47  427
1         93  30  217
```



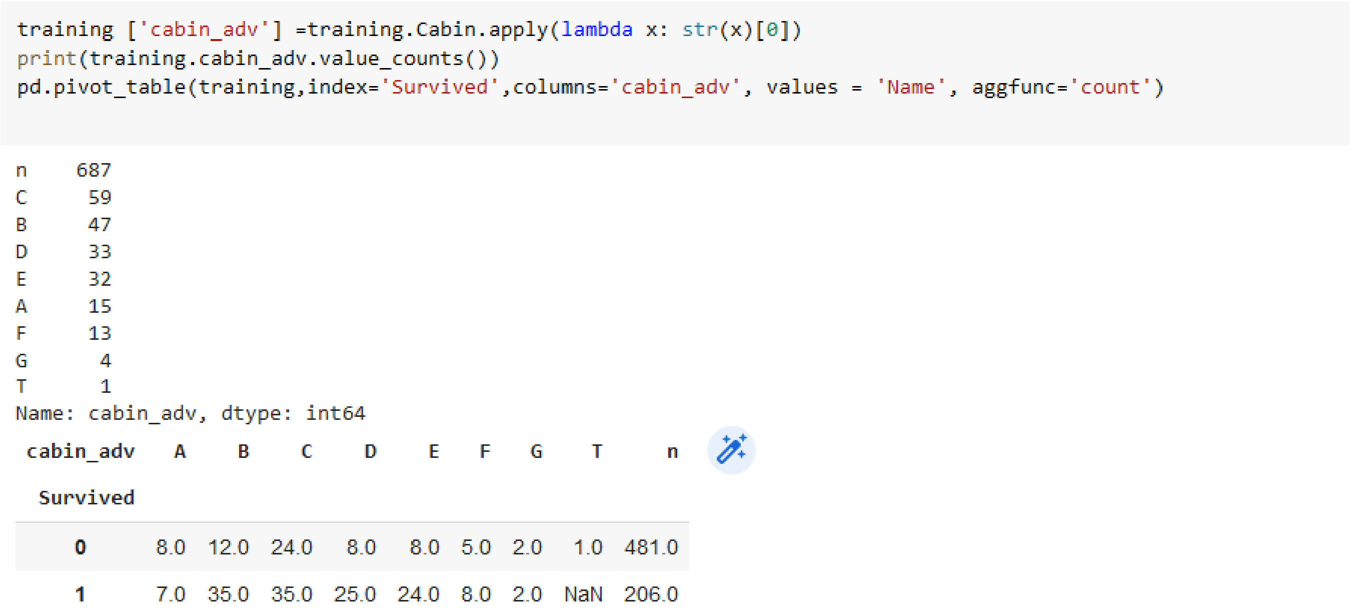
```
df_cat.Cabin
training['cabin_multiple'] = training.Cabin.apply(lambda x: 0 if pd.isna(x)
else len(x.split(' ')))
training['cabin_multiple'].value_counts()
```

```
0    687
1    180
2     16
3      6
4      2
Name: cabin_multiple, dtype: int64
```



```
pd.pivot_table(training, index = 'Survived', columns = 'cabin_multiple', values = 'Ticket' ,aggfunc ='count')
```

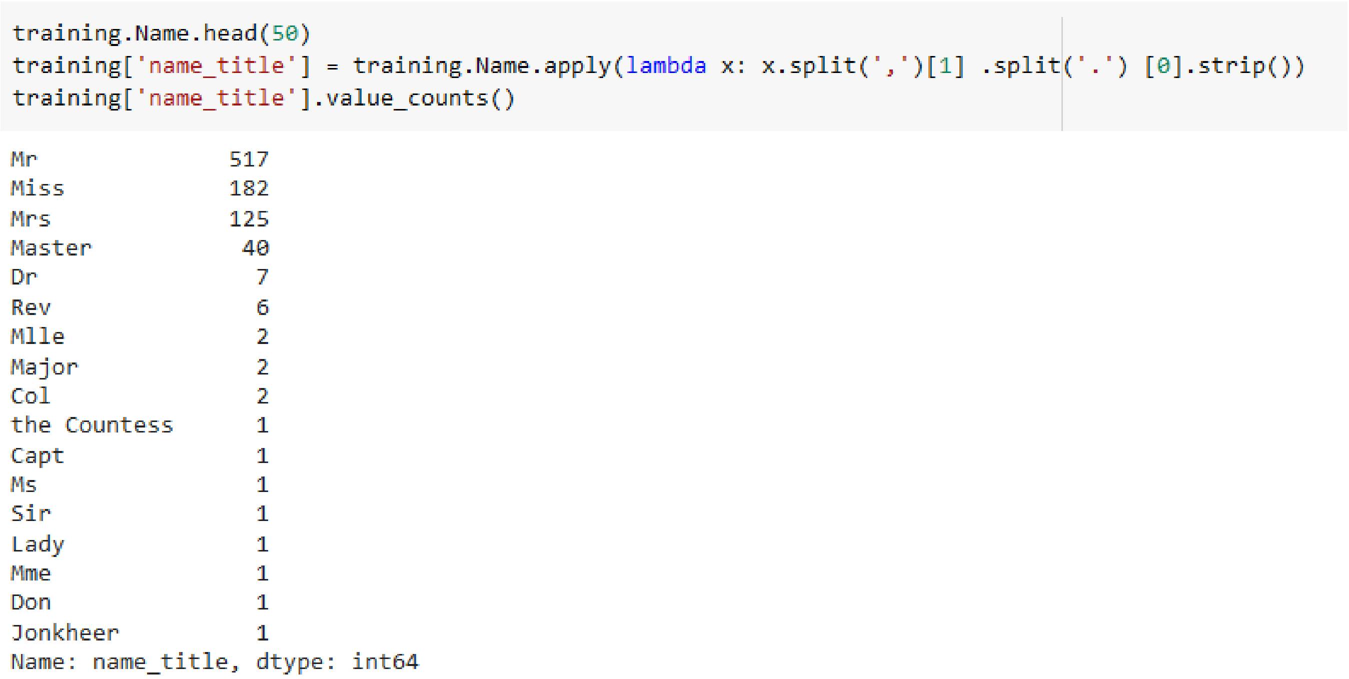
cabin_multiple	0	1	2	3	4
Survived					
0	481.0	58.0	7.0	3.0	NaN
1	206.0	122.0	9.0	3.0	2.0



```
training ['cabin_adv'] =training.Cabin.apply(lambda x: str(x)[0])
print(training.cabin_adv.value_counts())
pd.pivot_table(training,index='Survived',columns='cabin_adv', values = 'Name', aggfunc='count')
```

```
n    687
C     59
B     47
D     33
E     32
A     15
F     13
G      4
T      1
Name: cabin_adv, dtype: int64
```

cabin_adv	A	B	C	D	E	F	G	T	n
Survived									
0	8.0	12.0	24.0	8.0	8.0	5.0	2.0	1.0	481.0
1	7.0	35.0	35.0	25.0	24.0	8.0	2.0	NaN	206.0



```
training.Name.head(50)
training['name_title'] = training.Name.apply(lambda x: x.split(',')[1] .split('.') [0].strip())
training['name_title'].value_counts()
```

```
Mr              517
Miss            182
Mrs             125
Master           40
Dr                7
Rev               6
Mlle              2
Major             2
Col               2
the Countess      1
Capt              1
Ms                1
Sir               1
Lady              1
Mme               1
Don               1
Jonkheer          1
Name: name_title, dtype: int64
```