In [1]:

```
pip install kaggle
```

```
Requirement already satisfied: kaggle in c:\programdata\anaconda3\lib\site-pack
ages (1.5.12)
Requirement already satisfied: python-slugify in c:\programdata\anaconda3\lib\s
ite-packages (from kaggle) (5.0.2)
Requirement already satisfied: tqdm in c:\programdata\anaconda3\lib\site-packag
es (from kaggle) (4.62.3)
Requirement already satisfied: six>=1.10 in c:\programdata\anaconda3\lib\site-p
ackages (from kaggle) (1.16.0)
Requirement already satisfied: certifi in c:\programdata\anaconda3\lib\site-pac
kages (from kaggle) (2021.10.8)
Requirement already satisfied: urllib3 in c:\programdata\anaconda3\lib\site-pac
kages (from kaggle) (1.26.7)
Requirement already satisfied: python-dateutil in c:\programdata\anaconda3\lib
\site-packages (from kaggle) (2.8.2)
Requirement already satisfied: requests in c:\programdata\anaconda3\lib\site-pa
ckages (from kaggle) (2.26.0)
Requirement already satisfied: text-unidecode>=1.3 in c:\programdata\anaconda3
\lib\site-packages (from python-slugify->kaggle) (1.3)
Requirement already satisfied: charset-normalizer~=2.0.0 in c:\programdata\anac
onda3\lib\site-packages (from requests->kaggle) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in c:\programdata\anaconda3\lib\sit
e-packages (from requests->kaggle) (3.2)
Requirement already satisfied: colorama in c:\programdata\anaconda3\lib\site-pa
ckages (from tqdm->kaggle) (0.4.4)
Note: you may need to restart the kernel to use updated packages.
```

In [20]:

```
def extract_data(file_name, file_path):
    !kaggle competitions download titanic -f $file_name -p $file_path --force
```

In [21]:

```
import os
train_file_name="train.csv"
test_file_name="test.csv"

raw_data_path = os.path.join(os.path.pardir, "data", "raw")
extract_data(train_file_name, raw_data_path)
extract_data(test_file_name, raw_data_path)
```

```
Downloading train.csv to ..\data\raw



  0%|          | 0.00/59.8k [00:00<?, ?B/s]
100%|##########| 59.8k/59.8k [00:00<00:00, 1.58MB/s]

  0%|          | 0.00/28.0k [00:00<?, ?B/s]
100%|##########| 28.0k/28.0k [00:00<00:00, 2.18MB/s]

Downloading test.csv to ..\data\raw
```

In [29]:

```
import pandas as pd
```

In [39]:
```
data = pd.read_csv("train.csv")
```

In [40]:
```
data
```

Out[40]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Ca
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	
...	...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	

891 rows × 12 columns

In [43]:
```
import numpy as np
```

```
In [44]: data.head()
```

Out[44]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Na
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C8
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Na
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C12
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Na

```
In [48]: men = data.loc[data.Sex == "male"]["Survived"]
         rate_men = sum(men)/len(men)
```

```
In [50]: print("Percentage of men who survived:", rate_men)
```

```
Percentage of men who survived: 0.18890814558058924
```

In [52]:
```
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
```

In [53]:
```
data.dtypes
```

Out[53]:
```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

In [54]:
```
data.columns
```

Out[54]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

In [59]:
```
data.info(verbose=True, show_counts=True)
```

```
<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
```

In [60]:
```
data.describe()
```

Out[60]:

	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

In [63]:
```
data.describe(include="object")
```

Out[63]:

	Name	Sex	Ticket	Cabin	Embarked
count	891	891	891	204	889
unique	891	2	681	147	3
top	Braund, Mr. Owen Harris	male	347082	B96 B98	S
freq	1	577	7	4	644

In [64]:
```
data.isnull().sum(axis=0).sort_values(ascending=False)
```

Out[64]:
```
Cabin          687
Age            177
Embarked         2
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
SibSp            0
Parch            0
Ticket           0
Fare             0
dtype: int64
```

In [ ]:

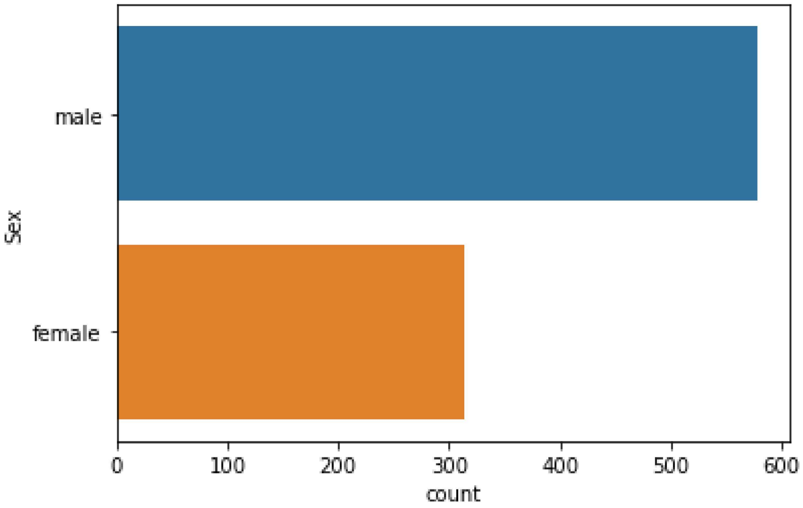
In [ ]:

In [ ]:

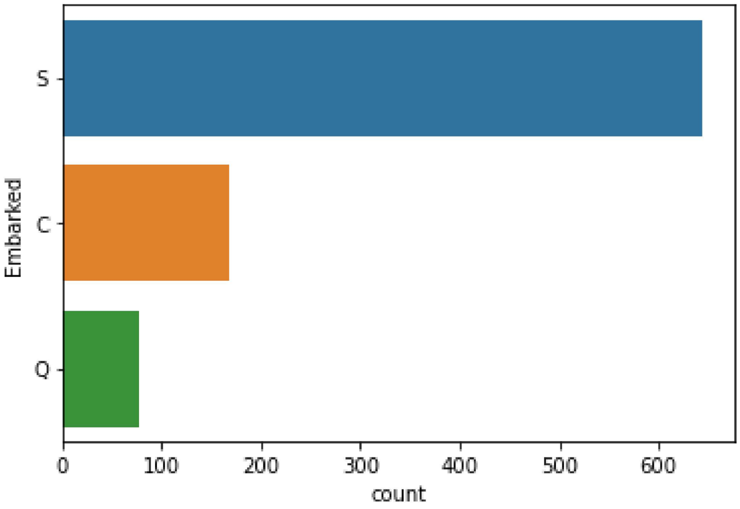
In [ ]:

In [ ]:

In [71]:



Sex
male
female
0
100
200
300
400
500
600
count



Embarked
S
C
Q
0
100
200
300
400
500
600
count

In [ ]: