In [17]:

```
import pandas as pd
df = pd.read_csv("train.csv")
```

In [18]:

```
#Print first 5 rows
df.head()
```

Out[18]:

	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

In [19]:

```
#Print last 5 rows
df.tail()
```

Out[19]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN	Q

In [20]:

```
#Obtain data on the different columns of the dataset
df.describe()
```

Out[20]:

	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 [23]:

```
#Sort dataset in alphabetical order of Name
#na_position helps to set a position for NaN values in the column 'Name'
SortNM = df.sort_values("Name", ascending=True, na_position='last')
SortNM.head()
```

Out[23]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
845	846	0	3	Abbing, Mr. Anthony	male	42.0	0	0	C.A. 5547	7.55	NaN	S
746	747	0	3	Abbott, Mr. Rossmore Edward	male	16.0	1	1	C.A. 2673	20.25	NaN	S
279	280	1	3	Abbott, Mrs. Stanton (Rosa Hunt)	female	35.0	1	1	C.A. 2673	20.25	NaN	S
308	309	0	2	Abelson, Mr. Samuel	male	30.0	1	0	P/PP 3381	24.00	NaN	C
874	875	1	2	Abelson, Mrs. Samuel (Hannah Wizosky)	female	28.0	1	0	P/PP 3381	24.00	NaN	C

In [28]:

```
#Check count of passengers
df["PassengerId"].shape[0]
```

Out[28]:

```
891
```

In [40]:

```
#Check count of different values in column 'Pclass'
df["Pclass"].value_counts()
```

Out[40]:

```
3    491
1    216
2    184
Name: Pclass, dtype: int64
```

In [41]:

```
#Check count of different values in column 'Survived'
df["Survived"].value_counts()
```

Out[41]:

```
0    549
1    342
Name: Survived, dtype: int64
```

In [27]:

```
#Check count of different values in column 'Embarked'
df["Embarked"].value_counts()
```

Out[27]:

```
S    644
C    168
Q     77
Name: Embarked, dtype: int64
```

In [31]:

```
#Check unique number of tickets out of 891 passengers
df["Ticket"].nunique()
```

Out[31]:

```
681
```

In [36]:

```
#Check unique number of Passenger classes and Sex
df[["Pclass","Sex"]].nunique()
```

Out[36]:

```
Pclass    3
Sex       2
dtype: int64
```

In [38]:

```
#Shows records where Embarked status is 'S'
EmbS = df["Embarked"]=='S'
df[EmbS].head()
```

Out[38]:

	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
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
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S

In [39]:

```
#Number of Female passengers
Fem = df["Sex"]=="female"
df[Fem].shape[0]
```

Out[39]:

```
314
```

In [49]:

```
#Number of passengers who Survived
Surv = df["Survived"]==1
df[Surv].shape[0]
```

Out[49]:

```
342
```

In [60]:

```
#Number of female passengers who survived
df[Fem & Surv].shape[0]
```

Out[60]:

```
233
```

In [48]:

```
#Number of passengers with Fare>100 (High Fare)
HF = df["Fare"]>100
df[HF].shape[0]
```

Out[48]:

```
53
```

In [61]:

```
#Number of passengers with Fare>100 who survived
df[HF & Surv].shape[0]
```

In [70]:

```
#Number of female passengers with Fare>100
df[Fem & HF].shape[0]
```

Out[70]:

```
34
```

In [66]:

```
#Number of female passengers with Fare>100 who survived
df[HF & Surv & Fem].shape[0]
```

Out[66]:

```
32
```

In [55]:

```
#Number of passengers with Fare<50 (Low Fare)
LF = df["Fare"]<50
df[LF].shape[0]
```

Out[55]:

```
730
```

In [62]:

```
#Number of passengers with Fare<50 who survived
df[LF & Surv].shape[0]
```

Out[62]:

```
233
```

In [64]:

```
#Survival rate (in %), among female passengers
FemSurv = df[Fem & Surv].shape[0]/df[Fem].shape[0] * 100
FemSurv
```

Out[64]:

```
74.20382165605095
```

In [65]:

```
#Survival rate (in %), among passengers with high fare
HFSurv = df[HF & Surv].shape[0]/df[HF].shape[0] * 100
HFSurv
```

Out[65]:

```
73.58490566037736
```

In [69]:

```
#Survival rate (as a % of ALL Female Passengers), among female passengers with high fare
FemHFSurv = df[HF & Surv & Fem].shape[0]/df[Fem & HF].shape[0] * 100
FemHFSurv
```

Out[69]:

```
94.11764705882352
```

In [72]:

```
#Survival rate (in %), among passengers with low fare
LFSurv = df[LF & Surv].shape[0]/df[LF].shape[0] * 100
LFSurv
```

Out[72]:

```
31.917808219178085
```