Titanic Dataset Analysis

Analyze the dataset to give probabilities of survival for each group. Each group is made based on age category, gender (sex), embarked (source of journey), and passenger class (pClass). The age categories are: [0-13], [13-18], [18, 40], [40, 100]. The embarked column represents: C → Cherbourg, Q → Queenstown. S → Southampton.

In [35]:
```
import pandas as pd
import numpy as np

# Load the dataset
df = pd.read_csv("titanic.csv")
df
```

Out[35]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cal
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	N
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	N
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	N
...	...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	N
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	E
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	N
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C1
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	N

891 rows × 12 columns

In [36]:
```
df1 = df.loc[:,["Age","Sex","Embarked", "Pclass","Survived"]]
df1
```

Out[36]:

	Age	Sex	Embarked	Pclass	Survived
0	22.0	male	S	3	0
1	38.0	female	C	1	1
2	26.0	female	S	3	1
3	35.0	female	S	1	1
4	35.0	male	S	3	0
...	...	...	...	...	...
886	27.0	male	S	2	0
887	19.0	female	S	1	1
888	NaN	female	S	3	0
889	26.0	male	C	1	1
890	32.0	male	Q	3	0

891 rows × 5 columns

In [37]:
```
df1["Embarked"] = df1['Embarked'].replace({
    "C": "Cherbourg",
    "Q": "Queenstown",
    "S": "Southamton"
})
df1
```

Out[37]:

	Age	Sex	Embarked	Pclass	Survived
0	22.0	male	Southamton	3	0
1	38.0	female	Cherbourg	1	1
2	26.0	female	Southamton	3	1
3	35.0	female	Southamton	1	1
4	35.0	male	Southamton	3	0
...	...	...	...	...	...
886	27.0	male	Southamton	2	0
887	19.0	female	Southamton	1	1
888	NaN	female	Southamton	3	0
889	26.0	male	Cherbourg	1	1
890	32.0	male	Queenstown	3	0

891 rows × 5 columns

In [38]:
```
# Define bins and labels
bins = [0, 13, 18, 40, 100]
labels = ["0-13", "13-18", "18-40", "40-100"]

# Categorize age groups
df1["Age_group"] = pd.cut(df1["Age"], bins=bins, labels=labels, right=True)
df1
```

Out[38]:

	Age	Sex	Embarked	Pclass	Survived	Age_group
0	22.0	male	Southamton	3	0	18-40
1	38.0	female	Cherbourg	1	1	18-40
2	26.0	female	Southamton	3	1	18-40
3	35.0	female	Southamton	1	1	18-40
4	35.0	male	Southamton	3	0	18-40
...	...	...	...	...	...	...
886	27.0	male	Southamton	2	0	18-40
887	19.0	female	Southamton	1	1	18-40
888	NaN	female	Southamton	3	0	NaN
889	26.0	male	Cherbourg	1	1	18-40
890	32.0	male	Queenstown	3	0	18-40

891 rows × 6 columns

In [39]:
```
# Group data by multiple factors and calculate survival probability
grouped = df1.groupby(["Age_group", "Sex", "Embarked", "Pclass"])["Survived"].mean(

# Round probabilities for readability
grouped["Survival_Probability"] = np.round(grouped["Survived"], 3)
grouped = grouped.drop(columns="Survived")
```

In [34]:
```
# Display results
print('\n--- Survival Probability by Group ---')
print(grouped.head(20))
```

```
--- Survival Probability by Group ---
   Age_group     Sex    Embarked  Pclass  Survival_Probability
0       0-13  female   Cherbourg       1                   NaN
1       0-13  female   Cherbourg       2                 1.000
2       0-13  female   Cherbourg       3                 0.857
3       0-13  female  Queenstown       1                   NaN
4       0-13  female  Queenstown       2                   NaN
5       0-13  female  Queenstown       3                   NaN
6       0-13  female  Southamton       1                 0.000
7       0-13  female  Southamton       2                 1.000
8       0-13  female  Southamton       3                 0.353
9       0-13    male   Cherbourg       1                   NaN
10      0-13    male   Cherbourg       2                 1.000
11      0-13    male   Cherbourg       3                 0.667
12      0-13    male  Queenstown       1                   NaN
13      0-13    male  Queenstown       2                   NaN
14      0-13    male  Queenstown       3                 0.000
15      0-13    male  Southamton       1                 1.000
16      0-13    male  Southamton       2                 1.000
17      0-13    male  Southamton       3                 0.389
18     13-18  female   Cherbourg       1                 1.000
19     13-18  female   Cherbourg       2                 1.000
```

In [ ]: