In [47]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [48]:
```
df=pd.read_csv(r"F:\FSDS\EDA-test\Titanic Dataset.csv")
```

In [49]:
```
df
```

Out[49]:

	sex	age	sibsp	parch	fare	embarked	class	who	alone	survived
0	male	22.0	1	0	7.2500	S	Third	man	False	0
1	female	38.0	1	0	71.2833	C	First	woman	False	1
2	female	26.0	0	0	7.9250	S	Third	woman	True	1
3	female	35.0	1	0	53.1000	S	First	woman	False	1
4	male	35.0	0	0	8.0500	S	Third	man	True	0
...	...	...	...	...	...	...	...	...	...	...
886	male	27.0	0	0	13.0000	S	Second	man	True	0
887	female	19.0	0	0	30.0000	S	First	woman	True	1
888	female	NaN	1	2	23.4500	S	Third	woman	False	0
889	male	26.0	0	0	30.0000	C	First	man	True	1
890	male	32.0	0	0	7.7500	Q	Third	man	True	0

891 rows × 10 columns

In [50]:
```
df.shape
```

Out[50]:
```
(891, 10)
```

1.Find the discrete and continuous variables in the dataset.

In [51]:
```
cat=df.select_dtypes(include='object').columns
cat
```

Out[51]:
```
Index(['sex', 'embarked', 'class', 'who'], dtype='object')
```

In [52]:
```
num=df.select_dtypes(exclude='object').columns
num
```

Out[52]:
```
Index(['age', 'sibsp', 'parch', 'fare', 'alone', 'survived'], dtype='object')
```

2.Check whether the dataset has missing values. If yes then treat them.

In [53]:
```
df.isnull().sum()
```

Out[53]:
```
sex           0
age         177
sibsp         0
parch         0
fare          0
embarked      2
class         0
who           0
alone         0
survived      0
dtype: int64
```

missing value treatment

In [54]:
```
# numerical column filled with median
# categorical column filled with mode
```

In [55]:
```
df.head()
```

Out[55]:

	sex	age	sibsp	parch	fare	embarked	class	who	alone	survived
0	male	22.0	1	0	7.2500	S	Third	man	False	0
1	female	38.0	1	0	71.2833	C	First	woman	False	1
2	female	26.0	0	0	7.9250	S	Third	woman	True	1
3	female	35.0	1	0	53.1000	S	First	woman	False	1
4	male	35.0	0	0	8.0500	S	Third	man	True	0

In [56]:
```
median=df['age'].median()
df['age'].fillna(median, inplace=True)
```

In [57]:
```
df['age'].isnull().sum()
```

Out[57]:
```
0
```

In [58]:
```
mode=df['embarked'].mode()[0]
df['embarked'].fillna(mode, inplace=True)
```

In [59]:
```
df['embarked'].isnull().sum()
```

Out[59]:
```
0
```

In [60]:
```
df.isnull().sum()
```

Out[60]:
```
sex         0
age         0
sibsp       0
parch       0
fare        0
embarked    0
class       0
who         0
alone       0
survived    0
dtype: int64
```

3.Difference between loc and iloc.Using loc function filter any 3 columns and iloc function filter any 2 columns with the index from 200 to 300

loc

In [61]:
```
df.loc[:, ['age', 'class', 'alone']]
```

Out[61]:

	age	class	alone
0	22.0	Third	False
1	38.0	First	False
2	26.0	Third	True
3	35.0	First	False
4	35.0	Third	True
...	...	...	...
886	27.0	Second	True
887	19.0	First	True
888	28.0	Third	False
889	26.0	First	True
890	32.0	Third	True

891 rows × 3 columns

iloc

In [71]:
```
df.iloc[200:300,[2,4,5]]
```

Out[71]:

	sibsp	fare	embarked
200	0	9.5000	S
201	8	69.5500	S
202	0	6.4958	S
203	0	7.2250	C
204	0	8.0500	S
...	...	...	...
295	0	27.7208	C
296	0	7.2292	C
297	1	151.5500	S
298	0	30.5000	S
299	0	247.5208	C

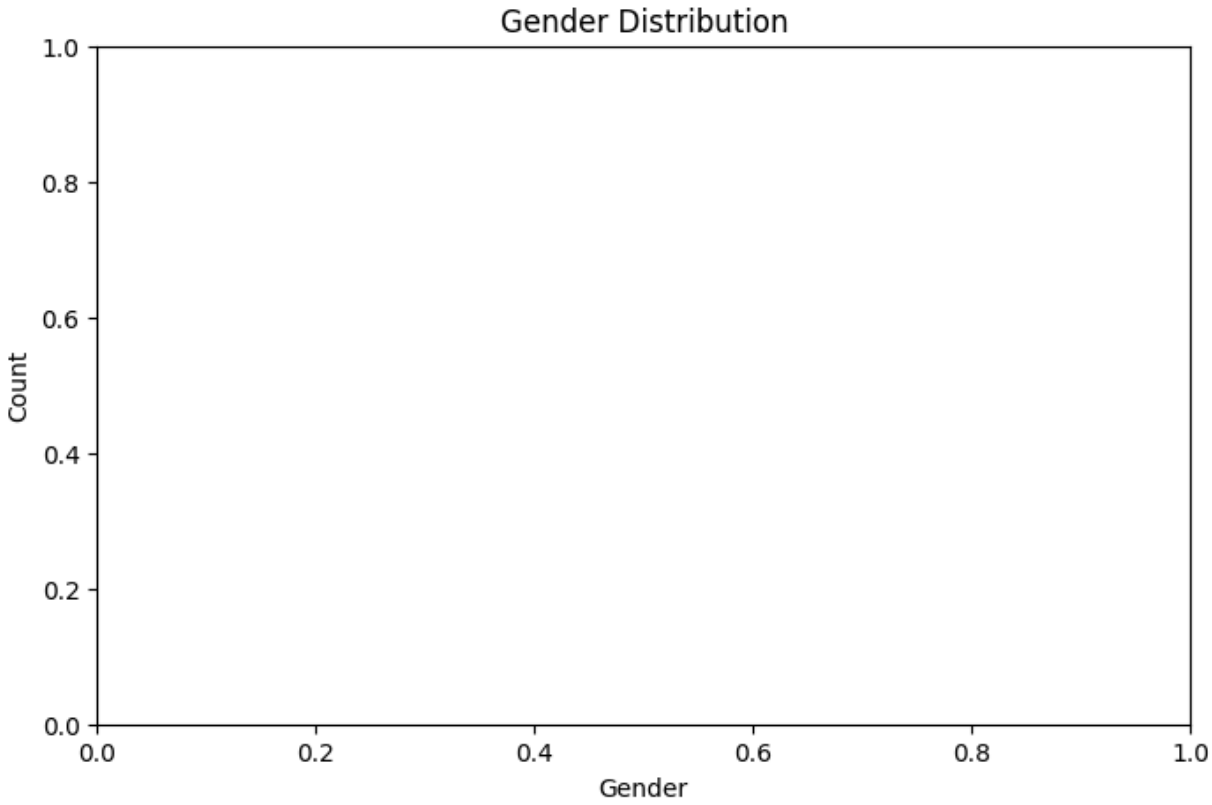
100 rows × 3 columns

4.Find the distribution of gender/sex column within the dataset using matplotlib and seaborn.

In [63]:

```
plt.figure(figsize=(8, 5))
#sns.countplot(x='Gender', data=df, palette='coolwarm')

plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```

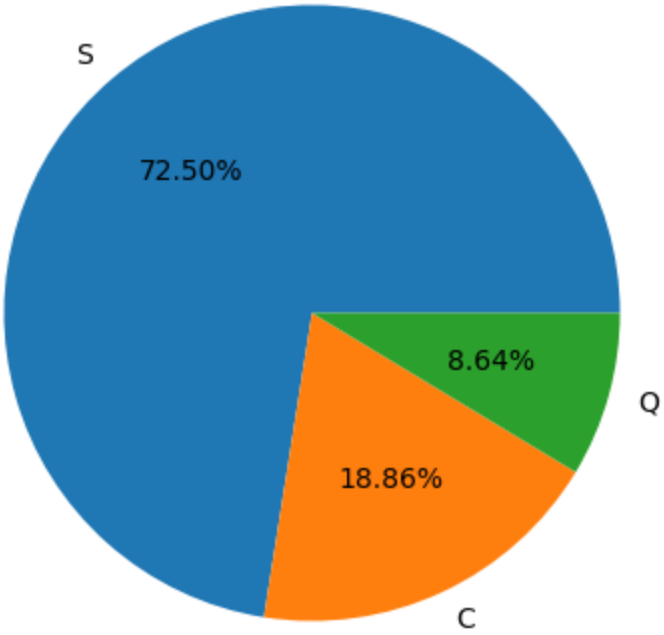


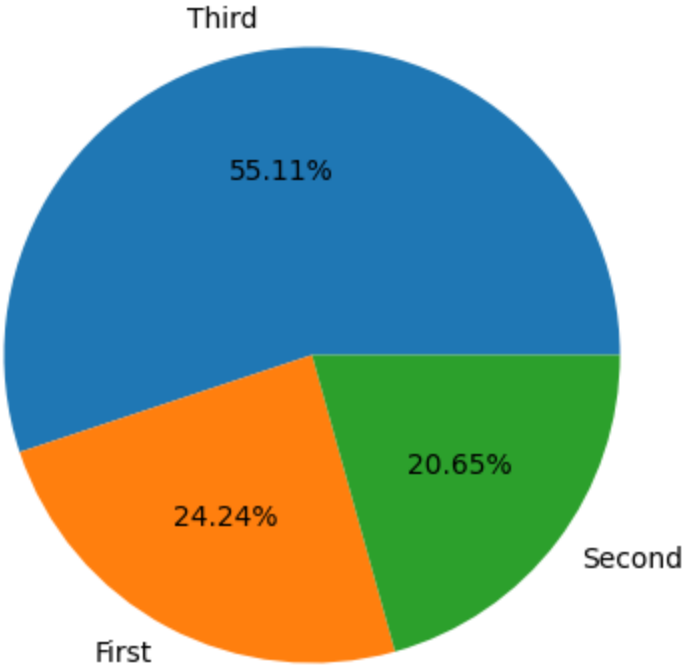
In [64]:
```
## 5.Plot pie chart of categorical columns.
```

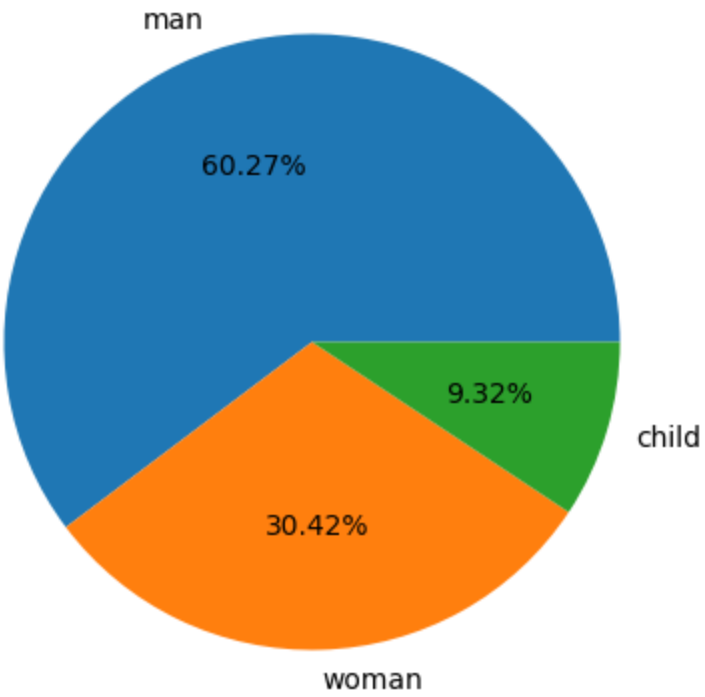
In [65]:
```
cat
```

Out[65]:
```
Index(['sex', 'embarked', 'class', 'who'], dtype='object')
```

In [69]:
```
for i in cat[1:]:
    df[i].value_counts()
    keys=df[i].value_counts().keys()
    values=df[i].value_counts().values
    plt.figure(figsize=(10,5))
    plt.pie(values,labels=keys,autopct='%0.2f%%')
```


S
72.50%
8.64%
Q
18.86%
C


Third
55.11%
20.65%
24.24%
Second
First



In [ ]:

In [ ]: