Task 2 :- Perform data cleaning and exploratory data analysis (EDA) on a dataset of your choice, such as the Titanic dataset from Kaggle. Explore the relationships between variables and identify patterns and trends in the data.

In [2]:
```
import pandas as pd
```

In [3]:
```
import matplotlib.pyplot as plt
```

In [4]:
```
df = pd.read_csv(r"C:\PYTHON DATASET\gender_submission.csv")
```

In [5]:
```
df.head()
```

Out[5]:

	PassengerId	Survived
0	892	0
1	893	1
2	894	0
3	895	0
4	896	1

In [6]:
```
print("SHAPE:",df.shape)
```

```
SHAPE: (418, 2)
```

In [7]:
```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  418 non-null    int64
 1   Survived     418 non-null    int64
dtypes: int64(2)
memory usage: 6.7 KB
None
```

In [8]:
```
print(df.isnull().sum())
```

```
PassengerId    0
Survived       0
dtype: int64
```

In [9]:
```
df.describe()
```

Out[9]:

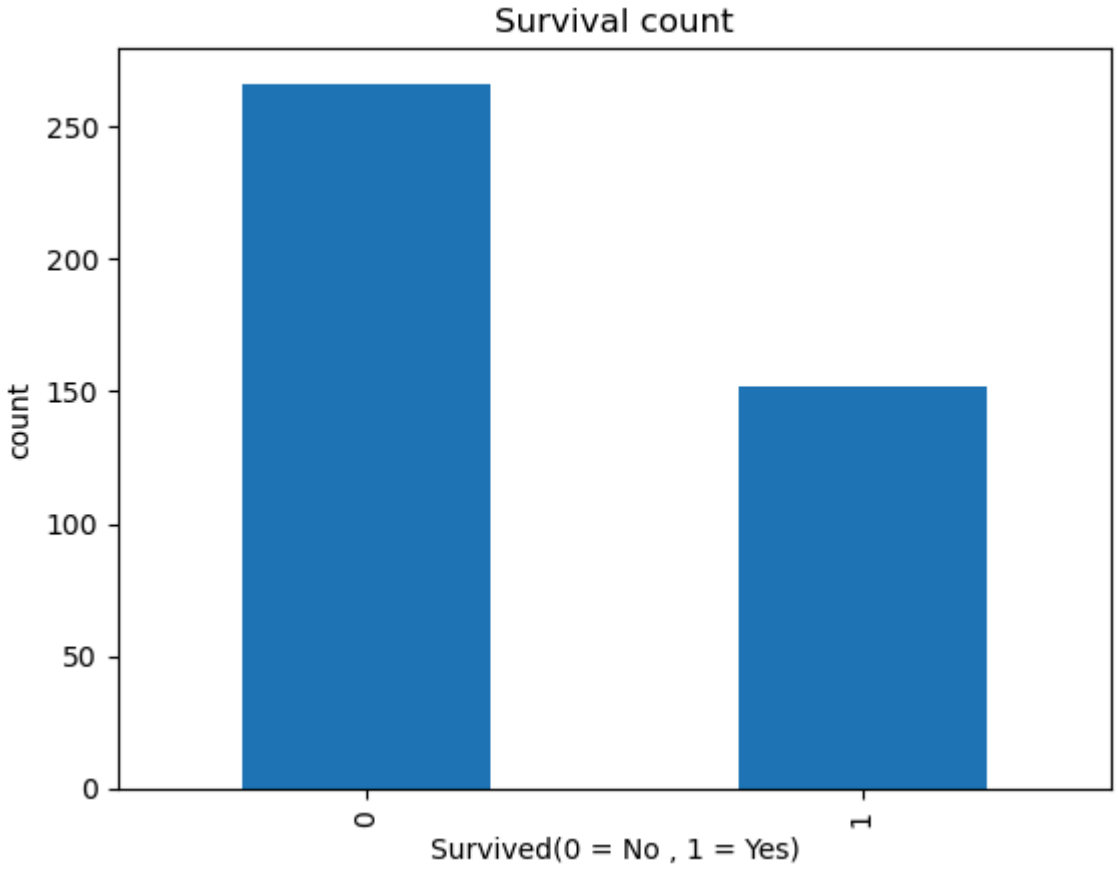
	PassengerId	Survived
count	418.000000	418.000000
mean	1100.500000	0.363636
std	120.810458	0.481622
min	892.000000	0.000000
25%	996.250000	0.000000
50%	1100.500000	0.000000
75%	1204.750000	1.000000
max	1309.000000	1.000000

In [10]:
```
# DATA CLEANING
print("Duplicates:",df.duplicated().sum())
```

```
Duplicates: 0
```

In [11]:
```
# EXPLORATORY DATA ANALYSIS(EDA)
# 1) BAR CHART

plt.figure()
df["Survived"].value_counts().plot(kind="bar")
plt.xlabel("Survived(0 = No , 1 = Yes)")
plt.ylabel("count")
plt.title("Survival count")
plt.show()
```

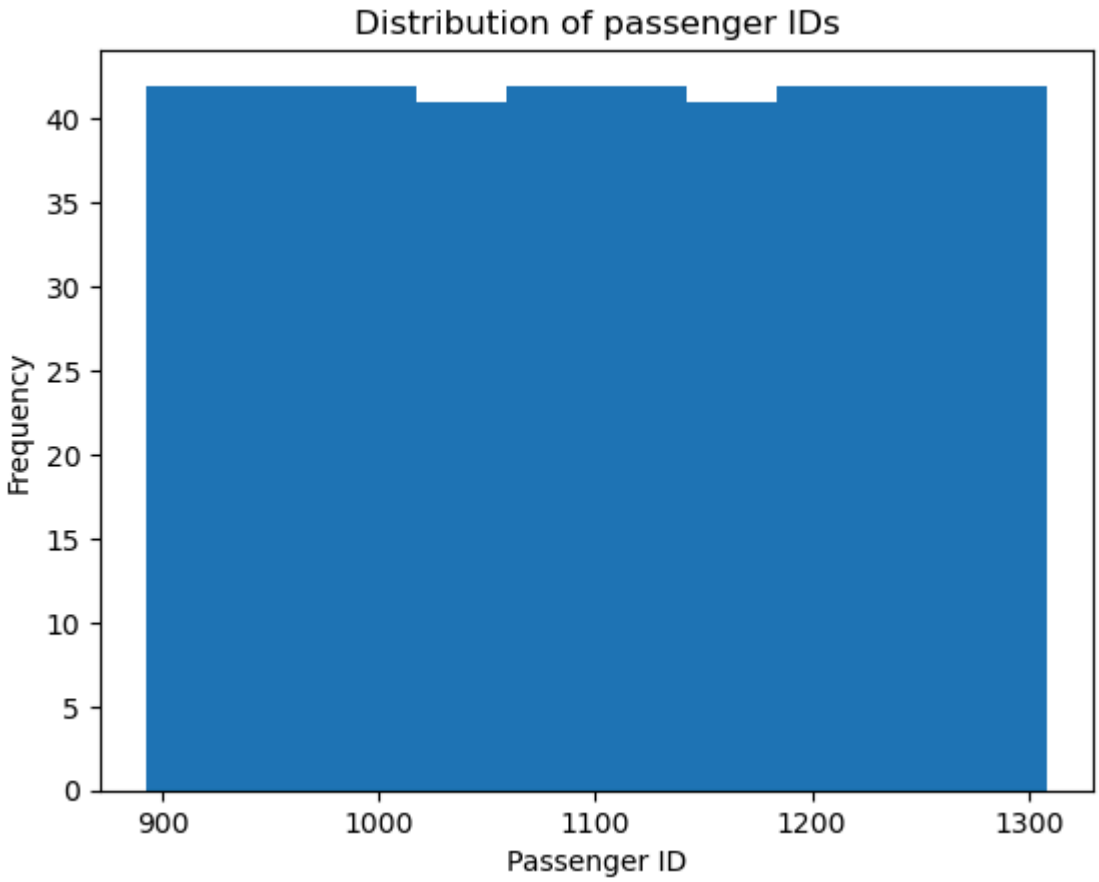


In [12]:
```
survival_rate = df["Survived"].value_counts(normalize=True)*100
print("Survival Percentage")
print(survival_rate)
```

```
Survival Percentage
0    63.636364
1    36.363636
Name: Survived, dtype: float64
```

In [13]:
```
# Distribution of Passenger IDs (Histogram)

plt.figure()
plt.hist(df["PassengerId"])
plt.xlabel("Passenger ID")
plt.ylabel("Frequency")
plt.title("Distribution of passenger IDs")
plt.show()
```



In [14]:
```
# CORRECTION ANALYSIS

print("Correlation Matrix:")
print(df.corr())
```

```
Correlation Matrix:
             PassengerId  Survived
PassengerId     1.000000 -0.023245
Survived       -0.023245  1.000000
```

Findings:

Total passengers: 418

Survival rate:

* Not Survived: ~63.6%

* Survived: ~36.4%

No missing values in this dataset.

PassengerId has no correlation with survival (as expected).

In [ ]: