

## Lab 1 Tasks

Q1. Load titanic Dataset from Kaggle and answer the following questions.

1. How many passengers were there on the Titanic?

```
import pandas as pd
data = pd.read_csv('Titanic-Dataset.csv')

total_passengers = len(data)
print("Total passengers:" + str(total_passengers))
```

⇒ Total passengers:891

2. What was the overall survival rate?

```
print(data['Survived'].mean() *100)
```

⇒ 38.38383838383838  
cell executed since last change

3. What was the average age of passengers?

```
print(data['Age'].mean() )
```

⇒ 29.69911764705882

4. How many male and female passengers were on board?

```
print(data['Sex'].value_counts() )
```

⇒ Sex  
male 577  
female 314  
Name: count, dtype: int64

5. What was the survival rate by gender?

```
GROUPS=data.groupby('Sex').agg(survival_rate=('Survived','mean'))
print(GROUPS*100)
```

	survival_rate
Sex	
female	74.203822
male	18.890815

6. How many passengers were in each class?

```
(data['Pclass'].value_counts() )
```

	count
Pclass	
3	491
1	216
2	184

7. What was the survival rate by class?

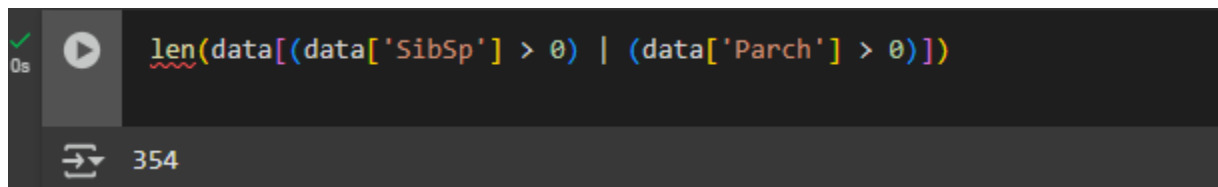
```
GROUPS=data.groupby('Pclass').agg(survival_rate=('Survived','mean'))
(GROUPS*100)
```

	survival_rate
Pclass	
1	62.962963
2	47.282609
3	24.236253

8. What was the average fare for each class?



9. How many passengers were traveling with family members (parents, siblings)?



10. How would you handle missing values in the dataset, particularly in the Age and Cabin columns?

For missing values in Age

```
data['Age'].fillna(data['Age'].median())
```

	Age
0	22.0
1	38.0
2	26.0
3	35.0
4	35.0
...	...
886	27.0
887	19.0
888	28.0
889	26.0
890	32.0

891 rows x 1 columns

dtype: float64

For missing values in Cabin

drop the column since many are NaN

```
data.drop('Cabin', axis=1, inplace=True)
```