Лабораторна робота № 2.
Аналіз даних про пасажирів лайнеру "Титанік"

В завданні пропонується за допомогою Pandas відповісти на декілька питань за даними репозиторія Kaggle (https://www.kaggle.com/c/titanic/data (https://www.kaggle.com/c/titanic/data)) (качати дані не потрібно – вони вже є в директорії роботи).

In [1]:
```
import numpy as np
import pandas as pd
%matplotlib inline
```

Зчитати дані з файлу в пам'ять у вигляді об'єкта Pandas.DataFrame

In [2]:
```
data = pd.read_csv('data/titanic_train.csv', index_col='PassengerId')
```

Дані представлені у вигляді таблиці. Подивимося на перші 5 рядків:

In [3]:
```
data.head(5)
```

Out[3]:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN

In [4]:
```
data.describe()
```

Out[4]:

	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Для прикладу відберемо пасажирів, які сіли в Cherbourg (Embarked=C) і заплатили більше 200$ за білет (fare > 200).

Переконайтеся, що Ви розумієте, як ця конструкція працює.
Якщо ні – подивіться як обчислюється вираз в квадратних дужках.

In [5]:
```
data[(data['Embarked'] == 'C') & (data.Fare > 200)].head()
```

Out[5]:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
119	0	1	Baxter, Mr. Quigg Edmond	male	24.0	0	1	PC 17558	247.5208	B58 B60
259	1	1	Ward, Miss. Anna	female	35.0	0	0	PC 17755	512.3292	NaN
300	1	1	Baxter, Mrs. James (Helene DeLaudeniere Chaput)	female	50.0	0	1	PC 17558	247.5208	B58 B60
312	1	1	Ryerson, Miss. Emily Borie	female	18.0	2	2	PC 17608	262.3750	B57 B59 B63 B66
378	0	1	Widener, Mr. Harry Elkins	male	27.0	0	2	113503	211.5000	C82

Можна відсортувати цих людей за зменшенням плати за білет.

In [6]:
```
data[(data['Embarked'] == 'C') &
     (data['Fare'] > 200)].sort_values(by='Fare',
                                       ascending=False).head()
```

Out[6]:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Em
259	1	1	Ward, Miss. Anna	female	35.0	0	0	PC 17755	512.3292	NaN	
680	1	1	Cardeza, Mr. Thomas Drake Martinez	male	36.0	0	1	PC 17755	512.3292	B51 B53 B55	
738	1	1	Lesurer, Mr. Gustave J	male	35.0	0	0	PC 17755	512.3292	B101	
312	1	1	Ryerson, Miss. Emily Borie	female	18.0	2	2	PC 17608	262.3750	B57 B59 B63 B66	
743	1	1	Ryerson, Miss. Susan Parker "Suzette"	female	21.0	2	2	PC 17608	262.3750	B57 B59 B63 B66	

Приклад створення ознаки.

In [7]:
```
def age_category(age):
    '''
    < 30 -> 1
    >= 30, <55 -> 2
    >= 55 -> 3
    '''
    if age < 30:
        return 1
    elif age < 55:
        return 2
    else:
        return 3
```

In [9]:
```
age_categories = [age_category(age) for age in data.Age]
```

In [10]:
```
data['Age_category'] = age_categories
```

Інший спосіб – через `apply`.

In [10]:
```
data['Age_category'] = data['Age'].apply(age_category)
```

1. Скільки чоловіків / жінок знаходилося на борту?

- 412 чоловіків і 479 жінок
- 314 чоловіків і 577 жінок
- 479 чоловіків і 412 жінок
- 577 чоловіків і 314 жінок

In [11]:
```
# Ваш код тут
```

2. Виведіть розподіл змінної `Pclass` (соціально-економічний статус) і цей же розподіл, тільки для чоловіків / жінок окремо. Скільки було чоловіків 2-го класу?

- 104
- 108
- 112
- 125

In [12]:
```
# Ваш код тут
```

3. Які значення медіани і стандартного відхилення платежів (`Fare`)? Виконайте округлення до 2 десяткових знаків.

- Медіана – 14.45, стандартне відхилення – 49.69
- Медіана – 15.1, стандартне відхилення – 12.15
- Медіана – 13.15, стандартне відхилення – 35.3
- Медіана – 17.43, стандартне відхилення – 39.1

In [13]:
```
# Ваш код тут
```

4. Чи правда, що люди молодші 30 років виживали частіше, ніж люди старші 60 років? Яка частка вижившних в обох групах?

- 22.7% серед молодих і 40.6% серед старих
- 40.6% серед молодих і 22.7% серед старих
- 35.3% серед молодих і 27.4% серед старих
- 27.4% серед молодих і 35.3% серед старих

In [14]:
```
# Ваш код тут
```

5. Чи правда, що жінки виживали частіше чоловіків? Яка частка вижившних в обох групах?

- 30.2% серед чоловіків і 46.2% серед жінок

- 35.7% серед чоловіків і 74.2% серед жінок
- 21.1% серед чоловіків і 46.2% серед жінок
- 18.9% серед чоловіків і 74.2% серед жінок

In [15]:
```
# Ваш код тут
```

6. Знайдіть найбільш популярні імена серед пасажирів Титаніку чоловічої статі

- Charles
- Thomas
- William
- John

In [16]:
```
# Ваш код тут
```

7. Порівняйте графічно розподіли вартості білетів і віку у врятованих та загиблих. Середній вік загиблих вище, правильно?

- Так
- Ні

In [12]:
```
# Ваш код тут
```

8. Як відрізняється середній вік чоловіків / жінок в залежності від класу обслуговування? Оберіть правильні твердження:

- В середньому чоловіки 1-го класу старші 40 років
- В середньому жінки 1-го класу старші 40 років
- Чоловіки всіх класів в середньому старші жінок того ж класу
- В середньому люди в 1 класі старші, ніж в 2-му, а також старші представників 3-го класу

In [11]:
```
# Ваш код тут
```