Тема 1. Первичный анализ данных с Pandas

Практическое задание. Анализ данных пассажиров "Титаника"

Заполните код в клетках (где написано "Ваш код здесь") и ответьте на вопросы в веб-форме.

In [59]:
```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
%matplotlib inline
```

Считаем данные из файла в память в виде объекта Pandas.DataFrame

In [60]:
```
data = pd.read_csv('titanic_train.csv',
                   index_col='PassengerId')
data.shape
```

Out[60]: (891, 11)

Данные представлены в виде таблицы. Посмотрим на первые 5 строк:

In [61]:
```
data.head()
```

Out[61]:

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

In [62]:
```
data.describe()
```

Out[62]:

	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 примера отберем пассажиров, которые сели в Cherbourg (Embarked=C) и заплатили более 200 у.е. за билет (fare > 200).

Убедитесь, что Вы понимаете, как эта конструкция работает.
Если нет – посмотрите, как вычисляется выражение в квадратных в скобках.

In [63]:
```
data[(data['Embarked'] == 'C') & (data.Fare > 200)].head(20)
```

Out[63]:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
119	0	1	Baxter, Mr. Quigg Edmond	male	24.0	0	1	PC 17558	247.5208	B58 B60	C
259	1	1	Ward, Miss. Anna	female	35.0	0	0	PC 17755	512.3292	NaN	C
300	1	1	Baxter, Mrs. James (Helene DeLaudeniere Chaput)	female	50.0	0	1	PC 17558	247.5208	B58 B60	C
312	1	1	Ryerson, Miss. Emily Borie	female	18.0	2	2	PC 17608	262.3750	B57 B59 B63 B66	C
378	0	1	Widener, Mr. Harry Elkins	male	27.0	0	2	113503	211.5000	C82	C
381	1	1	Bidois, Miss. Rosalie	female	42.0	0	0	PC 17757	227.5250	NaN	C

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
558	0	1	Robbins, Mr. Victor	male	NaN	0	0	PC 17757	227.5250	NaN	C
680	1	1	Cardeza, Mr. Thomas Drake Martinez	male	36.0	0	1	PC 17755	512.3292	B51 B53 B55	C
701	1	1	Astor, Mrs. John Jacob (Madeleine Talmadge Force)	female	18.0	1	0	PC 17757	227.5250	C62 C64	C
717	1	1	Endres, Miss. Caroline Louise	female	38.0	0	0	PC 17757	227.5250	C45	C
738	1	1	Lesurer, Mr. Gustave J	male	35.0	0	0	PC 17755	512.3292	B101	C
743	1	1	Ryerson, Miss. Susan Parker "Suzette"	female	21.0	2	2	PC 17608	262.3750	B57 B59 B63 B66	C

Можно отсортировать этих людей по убыванию платы за билет.

In [64]:
```
data[(data['Embarked'] == 'C') &
     (data['Fare'] > 200)].sort_values(by='Fare',
                                       ascending=False).head(20)
```

Out[64]:

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

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
119	0	1	Baxter, Mr. Quigg Edmond	male	24.0	0	1	PC 17558	247.5208	B58 B60	C
300	1	1	Baxter, Mrs. James (Helene DeLaudeniere Chaput)	female	50.0	0	1	PC 17558	247.5208	B58 B60	C
381	1	1	Bidois, Miss. Rosalie	female	42.0	0	0	PC 17757	227.5250	NaN	C
558	0	1	Robbins, Mr. Victor	male	NaN	0	0	PC 17757	227.5250	NaN	C
701	1	1	Astor, Mrs. John Jacob (Madeleine Talmadge Force)	female	18.0	1	0	PC 17757	227.5250	C62 C64	C
717	1	1	Endres, Miss. Caroline Louise	female	38.0	0	0	PC 17757	227.5250	C45	C
378	0	1	Widener, Mr. Harry Elkins	male	27.0	0	2	113503	211.5000	C82	C

Пример создания признака.

In [65]:
```
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 [66]:
```
age_categories = [age_category(age) for age in data.Age]
```

In [67]:
```
data['Age_category'] = age_categories
```

Другой способ – через `apply` .

In [68]:
```
data['Age_category'] = data['Age'].apply(age_category)
```

1. Сколько мужчин / женщин находилось на борту?

- 412 мужчин и 479 женщин
- 314 мужчин и 577 женщин
- 479 мужчин и 412 женщин
- 577 мужчин и 314 женщин

In [69]:
```
# Ваш код здесь

print('Количество мужчин/женщин, кот. находились на борту:')
data.Sex.value_counts()
```

```
Количество мужчин/женщин, кот. находились на борту:
```

Out[69]:
```
male      577
female    314
Name: Sex, dtype: int64
```

2. Выведите распределение переменной `Pclass` (социально-экономический статус) и это же распределение, только для мужчин / женщин по отдельности. Сколько было мужчин 2-го класса?

- 104
- 108
- 112
- 125

In [197…
```
# Ваш код здесь

tab = pd.crosstab(data['Pclass'], data['Sex'])
print('Количесво пассажиров (мужчин/женщин) и их распределение по классам:','\n', tab)

print('\n','Количество мужчин 2-го класса = 108')
```

```
Количесво пассажиров (мужчин/женщин) и их распределение по классам:
 Sex     female  male
Pclass
1           94   122
2           76   108
3          144   347
```

```
Количество мужчин 2-го класса = 108
```

3. Каковы медиана и стандартное отклонение платежей (`Fare`)? Округлите до 2 десятичных знаков.

- Медиана – 14.45, стандартное отклонение – 49.69
- Медиана – 15.1, стандартное отклонение – 12.15
- Медиана – 13.15, стандартное отклонение – 35.3
- Медиана – 17.43, стандартное отклонение – 39.1

In [71]:
```
# Ваш код здесь

print('Медиана по столбцу Fare =',round(data['Fare'].median(),2))
print('Стандартное отклонение по столбцу Fare=',round(np.std(data['Fare']), 2))
```

```
Медиана по столбцу Fare = 14.45
Стандартное отклонение по столбцу Fare= 49.67
```

4. Правда ли, что люди моложе 30 лет выживали чаще, чем люди старше 60 лет? Каковы доли выживших в обеих группах?

- 22.7% среди молодых и 40.6% среди старых
- 40.6% среди молодых и 22.7% среди старых
- 35.3% среди молодых и 27.4% среди старых
- 27.4% среди молодых и 35.3% среди старых

In [86]:
```
data['Age_category'] = age_categories # Добавляем столбец, где указан возрастная категория
data.head()
```

Out[86]:

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

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Age_category
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	2

In [94]:
```
# Ваш код здесь
data[['Age_category', 'Survived']].groupby(['Age_category'], as_index=False).mean()
```

Out[94]:

	Age_category	Survived
0	1	0.406250
1	2	0.420139
2	3	0.296804

Как видно из анализа, ***из первой группы пассажиров (моложе 30 лет) выжило 40,6 %, из третей группы пассажиров (старше 60 лет) выжило 29,68 %***

5. Правда ли, что женщины выживали чаще мужчин? Каковы доли выживших в обеих группах?

- 30.2% среди мужчин и 46.2% среди женщин
- 35.7% среди мужчин и 74.2% среди женщин
- 21.1% среди мужчин и 46.2% среди женщин
- 18.9% среди мужчин и 74.2% среди женщин

In [151…
```
data[['Sex', 'Survived']].groupby(['Sex'], as_index=False).mean()
```

Out[151…

	Sex	Survived
0	female	0.742038
1	male	0.188908

Да, женщин выжило больше, чем мужчин. Верный ответ:

18,9 % среди мужчин и 74,2 % среди женщин

6. Найдите самое популярное имя среди пассажиров Титаника мужского пола?

- Charles
- Thomas
- William
- John

In [225…
```
data_name = data[data.Sex == 'male']['Name']
C = []
for i in data_name:
  if '(' in i:
    if ')' in i.split('(')[1].split(' ')[0]:
      C.append(i.split('(')[1].split(' ')[0].split(')')[0])
    else:
      C.append(i.split('(')[1].split(' ')[0])

  else:
    C.append(i.split('. ')[1].split(' ')[0])

pd.DataFrame.from_dict(C)[0].value_counts()
```

Out[225…
```
William     35
John        23
George      13
Thomas      13
Charles     12
            ..
Orsen        1
Neal         1
Nicholas     1
Philemon     1
Henrik       1
Name: 0, Length: 291, dtype: int64
```

Самое популярное имя среди пассажиров Титаника мужского пола - William

7. Сравните графически распределение стоимости билетов и возраста у спасенных и у погибших. Средний возраст погибших выше, верно?

- Да
- Нет

In [162…
```
data_group = pd.DataFrame(data, columns = ['Age', 'Survived', 'Fare'])
data_group
```

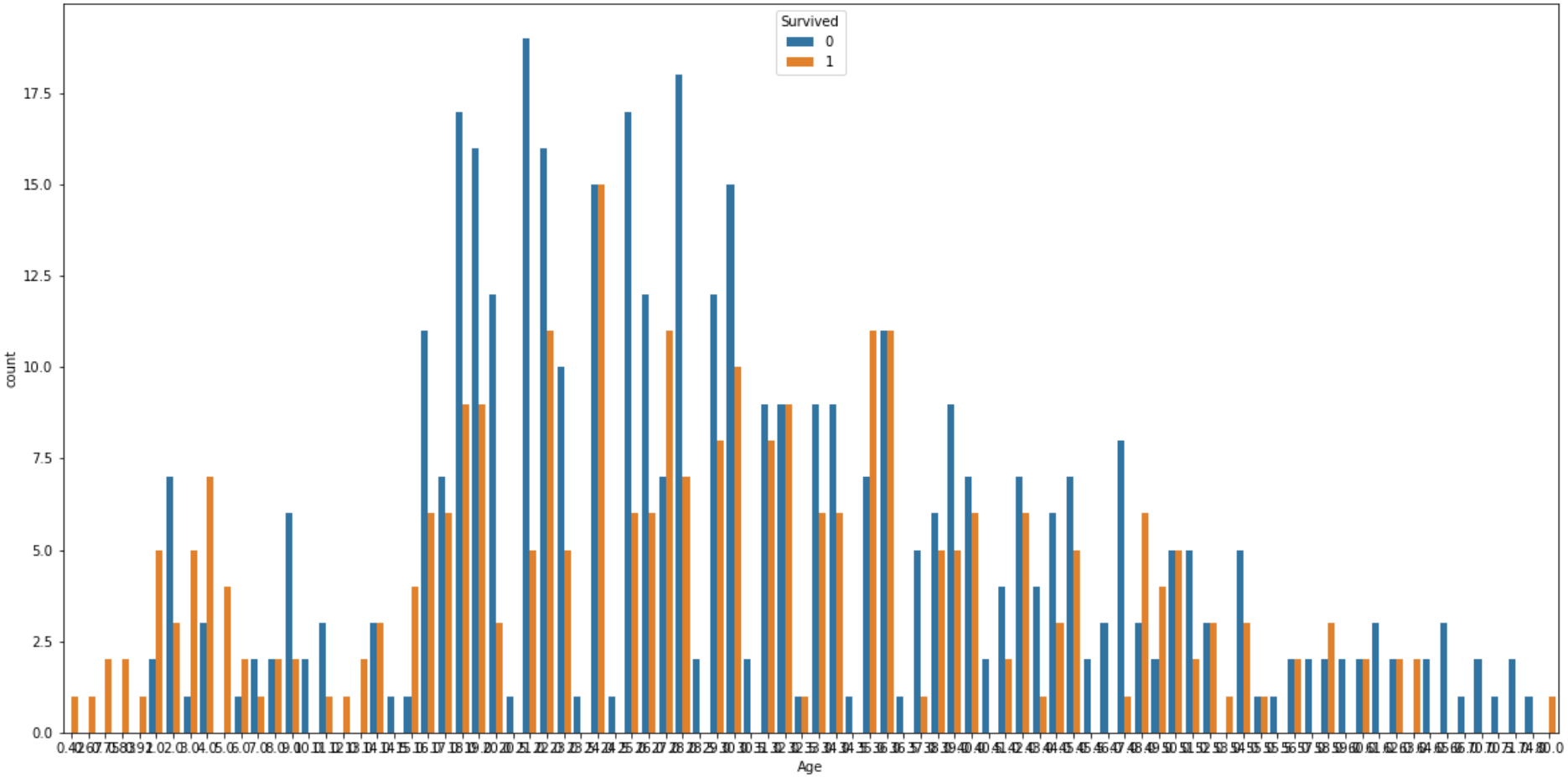
Out[162...

PassengerId	Age	Survived	Fare
1	22.0	0	7.2500
2	38.0	1	71.2833
3	26.0	1	7.9250
4	35.0	1	53.1000
5	35.0	0	8.0500
...	...	...	...
887	27.0	0	13.0000
888	19.0	1	30.0000
889	NaN	0	23.4500
890	26.0	1	30.0000
891	32.0	0	7.7500

891 rows × 3 columns

In [190...

```
plt.figure(figsize=(20,10))
sns.countplot(data=data_group, x=data_group.Age, hue='Survived')  # Строим диаграмму
plt.show()
```



Средний возраст погибших выше, чем у выживших верно?

ДА

8. Как отличается средний возраст мужчин / женщин в зависимости от класса обслуживания? Выберите верные утверждения:

- В среднем мужчины 1-го класса старше 40 лет
- В среднем женщины 1-го класса старше 40 лет
- Мужчины всех классов в среднем старше женщин того же класса
- В среднем люди в 1 классе старше, чем во 2-ом, а те старше представителей 3-го класса

In [150…

```
data_1 = data.groupby(['Sex', 'Pclass']).mean() # Группируем по нужным нам параметрам и считаем средний возраст
data_1
```

Out[150...

Sex	Pclass	Survived	Age	SibSp	Parch	Fare	Age_category
female	**1**	0.968085	34.611765	0.553191	0.457447	106.125798	1.851064
	2	0.921053	28.722973	0.486842	0.605263	21.970121	1.513158
	3	0.500000	21.750000	0.895833	0.798611	16.118810	1.770833
male	**1**	0.368852	41.281386	0.311475	0.278689	67.226127	2.131148
	2	0.157407	30.740707	0.342593	0.222222	19.741782	1.703704
	3	0.135447	26.507589	0.498559	0.224784	12.661633	1.812680

In [151...

```
pd.DataFrame(data_1['Age']) # Выводим нужный нам столбец
```

Out[151...

Sex	Pclass	Age
female	**1**	34.611765
	2	28.722973
	3	21.750000
male	**1**	41.281386
	2	30.740707
	3	26.507589

Верные утверждения:

- В среднем мужчины 1-го класса старше 40 лет

- Мужчины всех классов в среднем старше женщин того же класса

- В среднем люди в 1 классе старше, чем во 2-ом, а те старше представителей 3-го класса