Семинар: знакомство с Panda

B [1]:
```
import pandas as pd
import numpy as np
```

Задания для самостоятельного решения

Часть 1. Для датасета пассажиров Титаника

B [3]:
```
df = pd.read_csv('Data/titanic_train.csv", sep
df.head()
```

Out[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	Sil
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	
4	5	0	3	Allen, Mr. William Henry	male	35.0	

1. Какова доля семей, в которых минимальный возраст м
   детьми)?

B [4]:
```
df['Surname'] = df['Name'].apply(lambda name:
np.sum(df.groupby('Surname')['Age'].apply(min)
```

Out[4]:
```
0.17841079460269865
```

2. Какова доля выживших пассажиров класса 3? А пасса

В [5]:
```
print("Доля выживших третьего класса: ")
print(len(df[(df['Pclass'] == 3) & (df['Surviv
print("Доля выживших первого класса: ")
print(len(df[(df['Pclass'] == 1) & (df['Surviv
```

```
Доля выживших третьего класса
0.24236252545824846
Доля выживших первого класса
0.6296296296296297
```

3. Сколько пассажиров выжило, а сколько - нет?

В [6]:
```
print(np.sum([df['Survived'] == 1]))
print(np.sum([df['Survived'] == 0]))
```

```
342
549
```

4. Создайте столбец 'IsChild', который равен 1, если возр
   Для пропущенных значений поведение функции может

B [7]:

```
def fun(i):
    if i < 20:
        return 1
    else:
        return 0

df["IsChild"] = df["Age"].apply(fun)
df
```

Out[7]:

	PassengerId	Survived	Pclass	Name	Sex	Age
0	1	0	3	Braund, Mr. Owen Harris	male	22.0
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0
2	3	1	3	Heikkinen, Miss. Laina	female	26.0
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0
4	5	0	3	Allen, Mr. William Henry	male	35.0
...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN
889	890	1	1	Behr, Mr. Karl Howell	male	26.0
890	891	0	3	Dooley, Mr. Patrick	male	32.0

891 rows × 14 columns

5. Какова доля выживших женщин из первого класса? А из второго класса?

In [8]:
```
print('Доля выживших женщин из первого класса'
print(len(df[(df['Pclass'] == 1) & (df['Sex']
      / len(df[(df['Pclass'] == 1) & (df['Sex'
print('Доля выживших мужчин из третьего класса
print(len(df[(df['Pclass'] == 3) & (df['Sex']
      / len(df[(df['Pclass'] == 3) & (df['Sex'
```

```
Доля выживших женщин из первого класса
0.9680851063829787
Доля выживших мужчин из третьего класса
0.13544668587896252
```

Задания для самостоятельного решения

Часть 2.

Описание данных

В папке Data находится информация о студентах. Всего 10
делятся на две категории: * Students_info_i - информация с
Students_marks_i - оценки студентов из группы i за экзамен

Здесь можно посмотреть подробно о некоторых методах ра
данными pandas: https://www.kaggle.com/residentmario/rena
combining#Combining (https://www.kaggle.com/residentmario/
combining#Combining)

**Задача 2.1. Соберите всю информацию о студентах в о
получившейся таблице должна быть информация и оц
всех групп. Напечатайте несколько строк таблицы для
результата.¶**

B [9]:
```
df = pd.read_csv('Data/Students_info_0.csv').m
for i in range(1, 9):
    df_1 = pd.read_csv(f'Data/Students_info_{i
    df = pd.concat([df, df_1], ignore_index=Tr

df.head()
```

Out[9]:

	index	gender	race/ethnicity	parental level of education	lunch	prep
0	0	female	group B	bachelor's degree	standard	
1	1	female	group C	some college	standard	co
2	2	female	group B	master's degree	standard	
3	3	male	group A	associate's degree	free/reduced	
4	4	male	group C	some college	standard	

Задание 2.2. Удалите столбец index у полученной табли 10 строк таблицы.

B [10]:
```
df = df.drop(['index'], axis=1)
df.head(10)
```

Out[10]:

	gender	race/ethnicity	parental level of education	lunch	test preparation course
0	female	group B	bachelor's degree	standard	none
1	female	group C	some college	standard	completed
2	female	group B	master's degree	standard	none
3	male	group A	associate's degree	free/reduced	none
4	male	group C	some college	standard	none
5	female	group B	associate's degree	standard	none
6	female	group B	some college	standard	completed
7	male	group B	some college	free/reduced	none
8	male	group D	high school	free/reduced	completed
9	female	group B	high school	free/reduced	none

Задание 2.3. Выведите на экран размеры полученной т

In [11]:
```
1 df.shape
```

Out[11]:
```
(900, 9)
```

Задание 2.4. Выведите на экран статистические характ столбцов таблицы (минимум, максимум, среднее значе отклонение)

In [12]:
```
1 df.describe()
```

Out[12]:

	math score	reading score	writing score
count	900.000000	900.000000	900.000000
mean	66.045556	69.085556	67.910000
std	15.099269	14.568942	15.213194
min	0.000000	17.000000	10.000000
25%	57.000000	59.000000	57.000000
50%	66.000000	70.000000	69.000000
75%	77.000000	79.000000	79.000000
max	100.000000	100.000000	100.000000

Задание 2.5. Проверьте, есть ли в таблице пропущенны

In [13]:
```
1 df.isnull().sum()
```

Out[13]:
```
gender                         0
race/ethnicity                 0
parental level of education    0
lunch                          0
test preparation course        0
group                          0
math score                     0
reading score                  0
writing score                  0
dtype: int64
```

Задание 2.6. Выведите на экран средние баллы студен (math, reading, writing)

In [14]:
```
1 df.describe().iloc[1]
```

Out[14]:
```
math score       66.045556
reading score    69.085556
writing score    67.910000
Name: mean, dtype: float64
```

Задание 2.7. Как зависят оценки от того, проходил ли с подготовки к сдаче экзамена (test preparation course)? К

**каждого предмета в отдельности средний балл студен
для подготовки к экзамену и не проходивших курс.**

B [15]:
```
print("Средние баллы учеников, проходивших кур
df[df['test preparation course'] == 'completec
```

```
Средние баллы учеников, проходивших курс подготовк
```

Out[15]:
```
math score       69.814465
reading score    74.091195
writing score    74.613208
dtype: float64
```

B [16]:
```
print('Средние баллы учеников, не проходивших
df[df['test preparation course'] == 'none'].ag
```

```
Средние баллы учеников, не проходивших курс подгот
```

Out[16]:
```
math score       63.986254
reading score    66.350515
writing score    64.247423
dtype: float64
```

Задание 2.8. Выведите на экран все различные значени

B [17]:
```
[print(i) for i in df.lunch.unique()]
```

```
standard
free/reduced
```

Out[17]:
```
[None, None]
```

**Задание 2.9. Переименуйте колонку "parental level of ed
"test preparation course" в "test preparation" с помощью
rename**

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.rename.html (https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.rename.html)

B [18]:
```
df = df.rename(columns={'parental level of edu
df.head()
```

Out[18]:

	gender	race/ethnicity	education	lunch	test preparation
0	female	group B	bachelor's degree	standard	none
1	female	group C	some college	standard	completed
2	female	group B	master's degree	standard	none
3	male	group A	associate's degree	free/reduced	none
4	male	group C	some college	standard	none

Зафиксируем минимальный балл для сдачи экзамена

passmark = 50

Задание 2.10. Ответьте на вопросы:
* Какая доля студентов сдала экзамен по математике (pass
* Какая доля студентов, проходивших курс подготовки к экз
математике?
* Какая доля девушек, не проходивших курс подготовки к эк
по математике?

B [19]:
```
len(df[df['math score'] >= 50]) / len(df)
```

Out[19]: 0.8666666666666667

B [20]:
```
len(df[(df['math score'] >= 50) & (df['test pr
    / len(df[df['test preparation'] == 'comple
```

Out[20]: 0.9245283018867925

B [21]:
```
len(df[(df['math score'] >= 50) & (df['test pr
    / len(df[(df['test preparation'] == 'compl
```

Out[21]: 0.9156626506024096

**Задание 2.11. С помощью groupby выполните задания
время выполнения каждого из заданий.**
* Для каждой этнической группы выведите средний балл за
* Для каждого уровня образования выведите минимальный

B [22]:
```
df1 = df.groupby(['race/ethnicity']).agg({'rea
df1
```

Out[22]:

race/ethnicity	reading score
group A	65.194805
group B	67.359551
group C	69.037801
group D	69.643478
group E	73.056452

B [23]:
```
%%timeit
df.groupby(['race/ethnicity']).agg({'reading s
```

```
536 µs ± 10.4 µs per loop (mean ± std. dev. of 7 r
```

B [24]:
```
df2 = df.groupby(['education']).agg({'writing
df2
```

Out[24]:

education	writing score
associate's degree	35
bachelor's degree	38
high school	15
master's degree	46
some college	19
some high school	10

B [25]:
```
%%timeit
df.groupby(['education']).agg({'writing score'
```

```
594 µs ± 27.1 µs per loop (mean ± std. dev. of 7 r
```

Задание 2.12. Выполните задание 11 с помощью цикло выполнения.

B [26]:
```
df1_1 = pd.DataFrame({'race/ethnicity': df1.ir
s = df[['race/ethnicity', "reading score"]]
```

B [27]:
```
%%timeit
for i in df1_1.index:
    df1_1.loc[i] = (s[s['race/ethnicity'] ==
df1_1
```

```
1.33 ms ± 5.89 µs per loop (mean ± std. dev. of 7
```

B [28]:
```
df2_2 = pd.DataFrame({'education': df2.index,
s = df[['education', 'writing score']]
```

B [29]:
```
%%timeit
for i in df2_2.index:
    df2_2.loc[i] = min((s[s['education'] == i]
df2_2
```

```
1.38 ms ± 8.66 µs per loop (mean ± std. dev. of 7
```

Выполнение задания с помощью циклов занимает больше
этих же заданий с помощью groupby

**Задание 2.13. Выведите на экран средние баллы студе
предмету в зависимости от пола и уровня образовани
получиться количество групп, равных 2 * (число уровн
каждой такой группы выыведите средний балл по каж**

Это можно сделать с помощью сводных таблиц (pivot_table

https://www.kaggle.com/kamilpolak/tutorial-how-to-use-pivot-ta
(https://www.kaggle.com/kamilpolak/tutorial-how-to-use-pivot-t

B [30]:
```
df.groupby(['gender', 'education']).agg({'math
```

Out[30]:

gender	education	math score	reading score	writing
female	associate's degree	65.457944	74.093458	74.1
	bachelor's degree	68.232143	77.053571	78.2
	high school	61.012048	69.626506	68.2
	master's degree	64.468750	74.906250	75.9
	some college	65.446602	73.417476	73.8
	some high school	59.168675	69.192771	68.3
male	associate's degree	71.020202	67.797980	65.6
	bachelor's degree	70.431373	67.921569	67.2
	high school	63.806818	61.045455	57.7
	master's degree	74.476190	72.952381	71.8
	some college	68.479592	64.306122	62.4
	some high school	67.392405	64.189873	60.5

Задание 2.14. Сколько студентов успешно сдали экзам

Создайте новый столбец в таблице df под названием Math_
него F, если студент не сдал экзамен по математике (балл
иначе.

Посчитайте количество студентов, сдавших и не сдавших э

Сделайте аналогичные шаги для экзаменов по чтению и пи

B [31]:

```
df.loc[df['math score'] >= 50, 'Math_PassStatu
df.loc[df['math score'] < 50, 'Math_PassStatus

df.loc[df['reading score'] >= 50, 'Reading_Pas
df.loc[df['reading score'] < 50, 'Reading_Pass

df.loc[df['writing score'] >= 50, 'Writing_Pas
df.loc[df['writing score'] < 50, 'Writing_Pass

df
```

Out[31]:

	gender	race/ethnicity	education	lunch	te preparatic
0	female	group B	bachelor's degree	standard	nor
1	female	group C	some college	standard	complete
2	female	group B	master's degree	standard	nor
3	male	group A	associate's degree	free/reduced	nor
4	male	group C	some college	standard	nor
...	...	...	...	...	
895	female	group E	some high school	free/reduced	nor
896	male	group B	high school	free/reduced	nor
897	female	group B	some high school	free/reduced	complete
898	male	group D	associate's degree	standard	complete
899	female	group D	some high school	standard	complete

900 rows × 12 columns

Задание 2.15. Сколько студентов успешно сдали все эк

Создайте столбец OverAll_PassStatus и запишите в него дл
если студент не сдал хотя бы один из трех экзаменов, а ин

Посчитайте количество студентов, которые сдали все экза

B [36]:

```
1 df.loc[(df['math score'] >= 50) & (df['reading
2 df.loc[(df['math score'] < 50) | (df['reading
3 df
```

Out[36]:

	gender	race/ethnicity	education	lunch	te preparatic
0	female	group B	bachelor's degree	standard	nor
1	female	group C	some college	standard	complete
2	female	group B	master's degree	standard	nor
3	male	group A	associate's degree	free/reduced	nor
4	male	group C	some college	standard	nor
...	...	...	...	...	
895	female	group E	some high school	free/reduced	nor
896	male	group B	high school	free/reduced	nor
897	female	group B	some high school	free/reduced	complete
898	male	group D	associate's degree	standard	complete
899	female	group D	some high school	standard	complete

900 rows × 13 columns