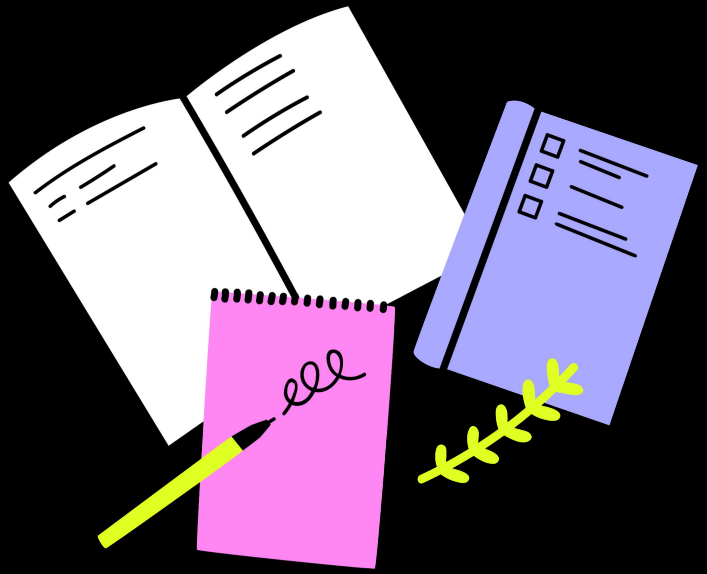


Семинар 3

Изменение таблиц в Pandas

Цели семинара №14:

- 📌 Научиться создавать, изменять и удалять признаки
- 📌 Изучить группировку данных и объединение таблиц
- 📌 Познакомиться со встроенными визуализациями

Викторина

Минутка самопроверки

Функция join в pandas объединяет по индексам?

1. Нет
2. Да

Функция join в pandas объединяет по индексам?

1. Нет
2. Да

Какой атрибут how стоял при объединение двух датафреймов с помощью merge по колонке col_1?

'LEFT DF'

	col_1	col_2
0	1	30
1	2	30
2	3	30
3	4	30

'RIGHT DF'

	col_1	col_3
0	1	40
1	3	50
2	5	50

'RESULT DF'

	col_1	col_2	col_3
0	1	30.0	40.0
1	2	30.0	NaN
2	3	30.0	50.0
3	4	30.0	NaN
4	5	NaN	50.0

1. Left
2. Right
3. Outer
4. Inner

<<0:40->>

Какой атрибут how стоял при объединение двух датафреймов с помощью merge по колонке col_1?

'LEFT DF'

	col_1	col_2
0	1	30
1	2	30
2	3	30
3	4	30

'RIGHT DF'

	col_1	col_3
0	1	40
1	3	50
2	5	50

'RESULT DF'

	col_1	col_2	col_3
0	1	30.0	40.0
1	2	30.0	NaN
2	3	30.0	50.0
3	4	30.0	NaN
4	5	NaN	50.0

1. Left
2. Right
3. Outer
4. Inner

По умолчанию метод .drop() возвращает новый датафрейм и не изменяет исходный?

1. Нет
2. Да

<<0:30->>

По умолчанию метод .drop() возвращает новый датафрейм и не изменяет исходный?

1. Нет
2. Да

Каким методом можно посчитать частотность появления уникальных значений в датафрейме/серии?

1. count_values()
2. value_counts()
3. count_unique()
4. value_unique()

<<0:30->>

Каким методом можно посчитать частотность появления уникальных значений в датафрейме/серии?

1. count_values()
2. value_counts()
3. count_unique()
4. value_unique()

Каким аргументом можно изменять исходный датафрейм, а не возвращать новый при использовании метода .drop()?

1. drop(keep=True)
2. drop(inplace=True)
3. drop(origin=True)
4. drop(new=False)

<<0:30->>

Каким аргументом можно изменять исходный датафрейм, а не возвращать новый при использовании метода .drop()?

1. drop(keep=True)
2. drop(inplace=True)
3. drop(origin=True)
4. drop(new=False)

Функция merge в pandas объединяет по индексам?

1. Нет
2. Да

<<0:20->>

Функция merge в pandas объединяет по индексам?

1. Нет
2. Да

Какой атрибут how стоял при объединение двух датафреймов с помощью merge по колонке col_1?

'LEFT DF'

	col_1	col_2
0	1	30
1	2	30
2	3	30
3	4	30

'RIGHT DF'

	col_1	col_3
0	1	40
1	3	50
2	5	50

'RESULT DF'

	col_1	col_2	col_3
0	1	30	40.0
1	2	30	NaN
2	3	30	50.0
3	4	30	NaN

1. Left
2. Right
3. Outer
4. Inner

<<0:40->>

Какой атрибут how стоял при объединение двух датафреймов с помощью merge по колонке col_1?

'LEFT DF'

	col_1	col_2
0	1	30
1	2	30
2	3	30
3	4	30

'RIGHT DF'

	col_1	col_3
0	1	40
1	3	50
2	5	50

'RESULT DF'

	col_1	col_2	col_3
0	1	30	40.0
1	2	30	NaN
2	3	30	50.0
3	4	30	NaN

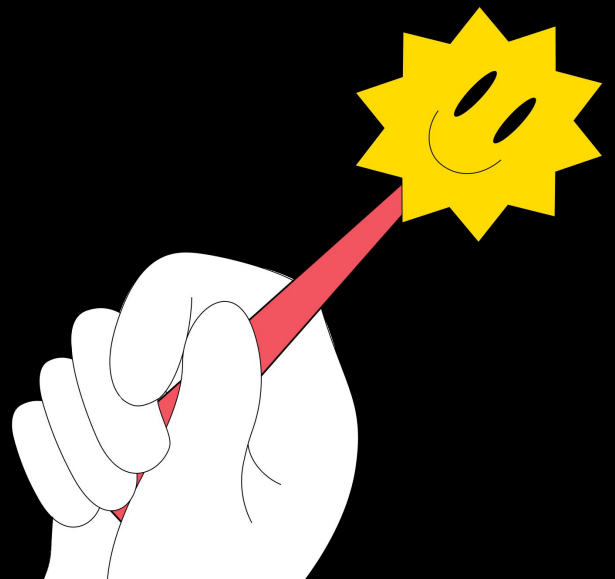
1. Left
2. Right
3. Outer
4. Inner

Ваши вопросы?

Изменение таблиц в Pandas

Практика

Изменение таблиц в Pandas



Задание 1.

1. Скачать данные по ссылке https://www.kaggle.com/datasets/ionaskel/laptop-prices
2. Считать данные с помощью pandas
3. Вывести на экран первые 5 строк

1.1 Создать новый признак Cpu_Company, который будет содержать только название фирмы, которая произвела CPU

1.2 Создать новый признак Memory_Amount, который будет содержать только количество Gb памяти без указания типа носителя

1.3 Создать новый признак Memory_Type, который будет содержать только тип носителя (HDD/SDD/др.)

1.4 Удалите признаки Memory и ScreenResolution

Задание 1.

1. Скачать данные по ссылке https://www.kaggle.com/datasets/ionaskel/laptop-prices
2. Считать данные с помощью pandas
3. Вывести на экран первые 5 строк

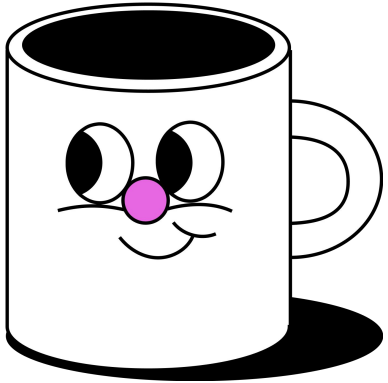
1.1 Создать новый признак Cpu_Company, который будет содержать только название фирмы, которая произвела CPU

1.2 Создать новый признак Memory_Amount, который будет содержать только количество Gb памяти без указания типа носителя

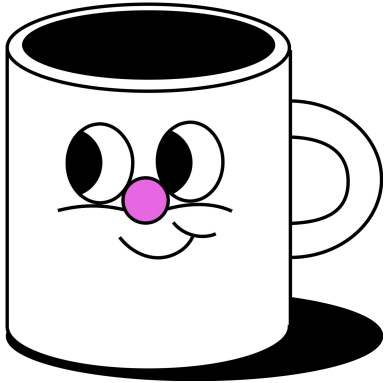
1.3 Создать новый признак Memory_Type, который будет содержать только тип носителя (HDD/SDD/др.)

1.4 Удалите признаки Memory и ScreenResolution

Перерыв



Перерыв



<<5:00->>

Задание 2.

2.1 Создайте признак SSD, который изначально равен 0

2.2 Поставьте в признаке SSD 1, если ноутбук действительно с типом носителя SSD

2.3 Уберите в признаке Weight значения 'kg' и поменяйте его тип данных на вещественный

Задание 2.

2.1 Создайте признак SSD, который изначально равен 0

2.2 Поставьте в признаке SSD 1, если ноутбук действительно с типом носителя SSD

2.3 Уберите в признаке Weight значения 'kg' и поменяйте его тип данных на вещественный

Задание 3.

Создайте датафрейм с клиентами:

```
clients = pd.DataFrame({
    'client_id': [45, 32, 67, 33, 43],
    'laptop_id': [506, 398, 710, 120, 1999]
})
```

laptop_id - это индексы датафрейма с ноутбуками

3.1 Присоедините к таблице clients данные по ноутбукам через метод join

3.2 Присоедините к таблице clients данные по ноутбукам через метод merge

Это нужно, чтобы понимать, какие ноутбуки покупались клиентами

Задание 3.

Создайте датафрейм с клиентами:

```
clients = pd.DataFrame({
    'client_id': [45, 32, 67, 33, 43],
    'laptop_id': [506, 398, 710, 120, 1999]
})
```

laptop_id - это индексы датафрейма с ноутбуками

3.1 Присоедините к таблице clients данные по ноутбукам через метод join

3.2 Присоедините к таблице clients данные по ноутбукам через метод merge

Это нужно, чтобы понимать, какие ноутбуки покупались клиентами

Задание 4.

Составьте несколько сводных таблиц

4.1 Найдите среднюю стоимость ноутбуков в зависимости от компании производителя

Отсортируйте от меньшей стоимости к большей

4.2 Найдите минимальную, среднюю и максимальную стоимости ноутбуков в зависимости от производителя процессора

4.3 Постройте таблицу с подсчетом количества ноутбуков в данных в зависимости от производителя CPU и ОЗУ

4.4 Постройте таблицу с подсчетом средней стоимости ноутбуков в данных в зависимости от операционной системы и GB памяти

Задание 4.

Составьте несколько сводных таблиц

4.1 Найдите среднюю стоимость ноутбуков в зависимости от компании производителя

Отсортируйте от меньшей стоимости к большей

4.2 Найдите минимальную, среднюю и максимальную стоимости ноутбуков в зависимости от производителя процессора

4.3 Постройте таблицу с подсчетом количества ноутбуков в данных в зависимости от производителя CPU и ОЗУ

4.4 Постройте таблицу с подсчетом средней стоимости ноутбуков в данных в зависимости от операционной системы и GB памяти

Задание 5*.

Ответьте на несколько вопросов

5.1 Ноутбуков каких компаний и с каким процессором больше?

5.2 С каким типом памяти и с каким объемом памяти больше ноутбуков?

Задание 5*.

Ответьте на несколько вопросов

5.1 Ноутбуков каких компаний и с каким процессором больше?

5.2 С каким типом памяти и с каким объемом памяти больше ноутбуков?

Ваши вопросы?

Подведем итоги

Домашнее задание

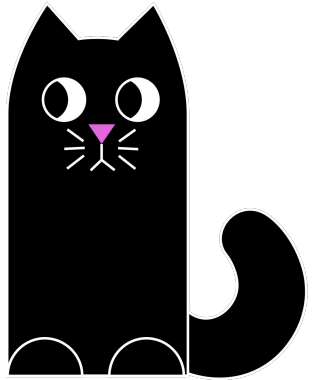
Домашнее задание 1

- Скачать данные по ссылке https://www.kaggle.com/datasets/ionaskel/laptop-prices
- Считать данные с помощью pandas
- Вывести на экран первые 5 строк

1.1 Создать новый признак price_per_sq_lot, который будет содержать среднюю стоимость за один кв. метр общей площади

1.2 Создать новый признак delta_renovated, который будет содержать разницу в годах между годом реновацией дома и годом постройки дома

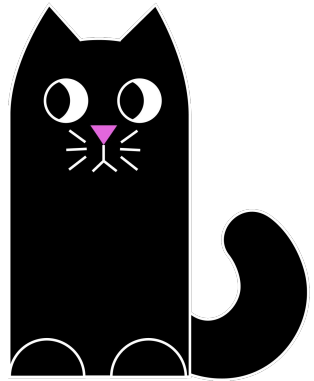
Если реновации дома не было, то в новом признаке поставьте 0



Домашнее задание 1

1.3 Создайте признаки года продажи, месяца продажи

1.4 Удалите признаки date, zipcode, lat, long



Домашнее задание 2

Создайте датафрейм с клиентами:

```
clients = pd.DataFrame({
    'client_id': [1459, 4684, 3498, 3942, 4535, 2308, 2866, 2765, 1472, 4236, 2295, 939, 3840,  280,   20, 4332, 3475, 4213, 3113,
4809, 2134, 2242, 2068, 4929, 1384, 1589, 3317, 2260, 1727, 1764, 1611, 1474],
    'house_id': [8965450190, 6823100225, 5104540330, 2131701075, 1522700060, 1189000207, 6821600300, 7137950720,
9510920050, 6131600255, 5428000070, 1788800910, 8100400160, 3123049142, 6306800010, 5083000375, 7920100025,
1951600150,  809001400,  339600110, 1622049154, 1099600250, 8563000110, 2768100205, 3995700435, 8861700030,
3303980210, 7731100066, 8146100580,  825069097, 3889100029, 9524100196]
})
```

house_id - это индексы датафрейма с домами

2.1 Присоедините к таблице clients данные по домам через метод join

2.2 Присоедините к таблице clients данные по домам через метод merge

Домашнее задание 3

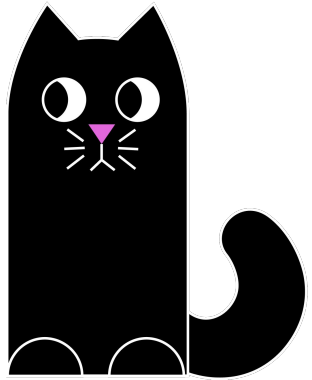
Составьте несколько сводных таблиц

3.1 Найдите среднюю стоимость домов в зависимости от количества спален

Отсортируйте от меньшей стоимости к большей

3.2 Найдите минимальную, среднюю и максимальную стоимости домов в зависимости от состояния дома

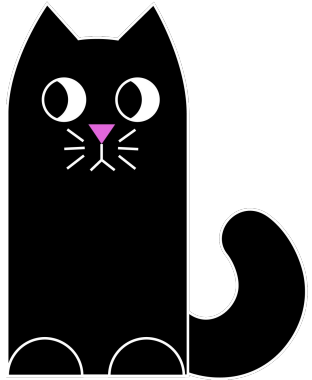
3.3 Постройте таблицу с подсчетом количества домов в данных в зависимости от вида на набережную и оценкой вида



Домашнее задание 3

3.4 Каких домов в зависимости от этажности и количества спален больше?

3.5 Постройте таблицу с подсчетом медианной стоимости домов в данных в зависимости от состояния дома и оценки дома



Спасибо за внимание