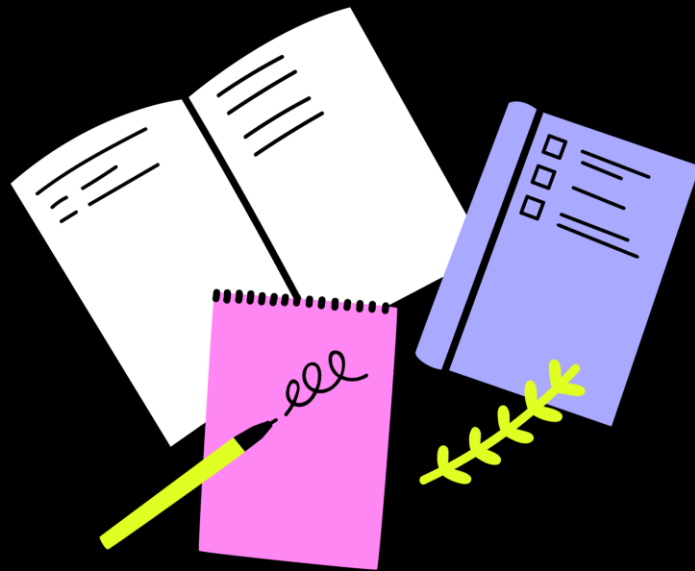




# Семинар 2

Анализ датасета с помощью Pandas



## Цели семинара №13:

- ☐ Узнать, как анализировать табличные данные
- ☐ Научиться считать статистики датафрейма
- ☐ Более детально изучить фильтрацию данных
- ☐ Разобраться с сортировками





# Викторина

Минутка самопроверки



## С помощью какого синтаксиса можно создать Pandas Series из Python списка?

1. `pd.MakeSeries(my_list)`
2. `pd.Series(my_list)`
3. `pd.GetSeries(my_list)`
4. `pd.CreateSeries(my_list)`

<<0:30->>

## С помощью какого синтаксиса можно создать Pandas Series из Python списка?

1. `pd.MakeSeries(my_list)`
2. `pd.Series(my_list)`
3. `pd.GetSeries(my_list)`
4. `pd.CreateSeries(my_list)`



## С помощью какого синтаксиса можно создать Pandas DataFrame?

1. `pd.MakeDataFrame(my_data)`
2. `pd.CreateDataFrame(my_data)`
3. `pd.GetDataFrame(my_data)`
4. `pd.DataFrame(my_data)`

<<0:30->>



## С помощью какого синтаксиса можно создать Pandas DataFrame?

1. `pd.MakeDataFrame(my_data)`
2. `pd.CreateDataFrame(my_data)`
3. `pd.dataframe(my_data)`
4. `pd.DataFrame(my_data)`



## Каким синтаксисом можно вернуть первые 20 строк из датафрейма?

1. `df.start(20)`
2. `df.head(20)`
3. `df.tail(20)`
4. `df.top(20)`

<<0:30->>



## Каким синтаксисом можно вернуть первые 20 строк из датафрейма?

1. `df.start(20)`
2. `df.head(20)`
3. `df.tail(20)`
4. `df.top(20)`



## Каким методом можно вернуть последние строки из датафрейма?

1. `df.last()`
2. `df.back()`
3. `df.tail()`
4. `df.bottom()`

<<0:30->>

## Каким методом можно вернуть последние строки из датафрейма?

1. `df.last()`
2. `df.back()`
3. `df.tail()`
4. `df.bottom()`



**Что из представленных статистик показывает самое частотное значение в данных?**

1. медиана
2. среднее
3. мода

<<0:30->>



## Что из представленных статистик показывает самое частотное значение в данных?

1. медиана
2. среднее
3. мода



**После данной сортировки первыми будут отображаться самые молодые клиенты `df.sort_values('Age', ascending=False)`?**

1. Правда
2. Ложь

<<0:30->>

После данной сортировки первыми будут отображаться самые молодые клиенты `df.sort_values('Age', ascending=False)`?

1. Правда
2. Ложь

```
df.sort_values('Age', ascending=False)
```

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age
<b>6443</b>	6444	15764927	Rogova	753	France	Male	92
<b>6759</b>	6760	15660878	T'ien	705	France	Male	92
<b>2458</b>	2459	15813303	Rearick	513	Spain	Male	88
<b>3033</b>	3034	15578006	Yao	787	France	Female	85
<b>3387</b>	3388	15798024	Lori	537	Germany	Male	84
...	...	...	...	...	...	...	...



## Какой будет вывод у данного кода?

```
sample_df.iloc[0]
```

```
sample_df
```

	Company	OpSys	TypeName
880	HP	Windows 10	2 in 1 Convertible
568	Lenovo	Windows 10	Notebook
634	Asus	Windows 10	Notebook
1134	Acer	Windows 10	Notebook
263	Dell	Windows 10	Notebook

1. Company HP  
OpSys Windows 10  
TypeName 2 in 1 Convertible
1. KeyError
2. IndexError

<<0:30->>



## Какой будет вывод у данного кода?

```
sample_df.iloc[0]
```

```
Company          HP  
OpSys            Windows 10  
TypeName    2 in 1 Convertible  
Name: 880, dtype: object
```

1. Company HP  
OpSys Windows 10  
TypeName 2 in 1 Convertible
1. KeyError
2. IndexError

```
sample_df
```

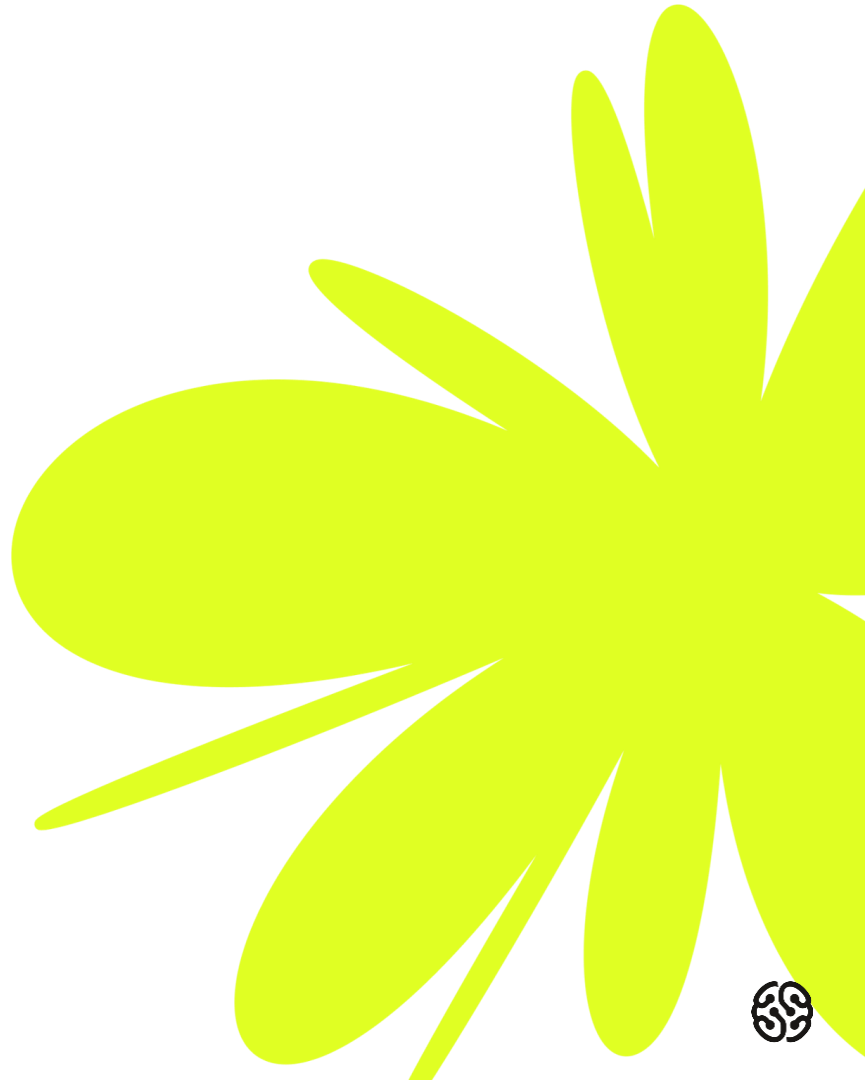
	Company	OpSys	TypeName
880	HP	Windows 10	2 in 1 Convertible
568	Lenovo	Windows 10	Notebook
634	Asus	Windows 10	Notebook
1134	Acer	Windows 10	Notebook
263	Dell	Windows 10	Notebook



Анализ датасета с помощью Pandas

# Ваши вопросы?

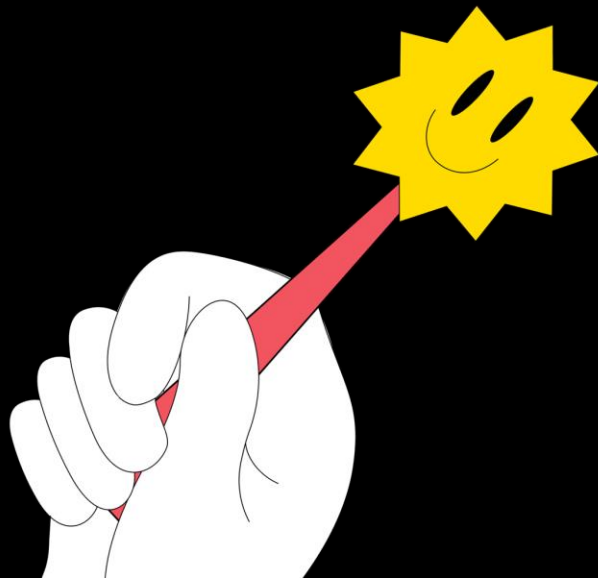
Анализ датасета с помощью Pandas





# Практика

Анализ датасета с помощью 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
4. Посмотреть на описание признаков и на их содержание



5 минут



## Задание 2.

Проведите первичный анализ данных

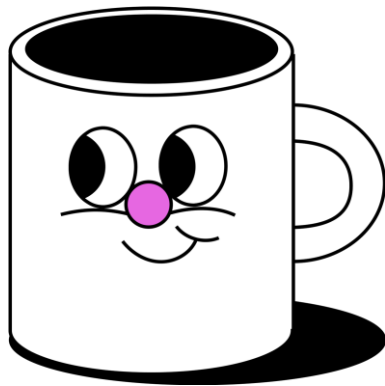
1. Изучите типы данных
2. Найдите количество пропущенных ячеек в данных
3. Посчитайте основные статистики по всем признакам и поизучайте их



**5 минут**



## Перерыв



## Задание 3.

Ответьте на несколько вопросов

**3.1 Ноутбуков от какой компании больше всего в наборе данных?**

**3.2 Какая минимальная и максимальная стоимость у ноутбуков в данных?**

**3.3 Какой самый дорогой ноутбук в данных?**

Выведите все характеристики только по этому ноутбуку

*Если таких ноутбуков несколько, то выводите их всех*



5 минут



## Задание 4.

Ответьте на несколько вопросов

### 4.1 Найдите ноутбуки с самой маленькой диагональю в данных?

Выведите только следующие характеристики:

1. Компания
2. Типа ноутбука
3. Диагональ
4. Стоимость

*Если таких ноутбуков несколько, то выводите их всех*

### 4.2 Сколько стоит самый дорогой ноутбук у компании HP?

### 4.3 Как много ноутбуков Ultrabook с 8GB RAM?

Найдите сколько таких ультрабуков с 8GB ОЗУ в процентном соотношении относительно всех ультрабуков



5 минут





## Задание 5.

### 5.1 Выберите ноутбук клиенту

Клиент хочет подобрать ноутбук с 8GB или 16GB ОЗУ на Windows 10 в стоимости до 500 евро, сколько у него вариантов?

### 5.2 Выберите ноутбук клиенту

Клиент хочет подобрать ноутбук от MSI, с видеокартой Nvidia GeForce GTX 1050 Ti и главное не с диагональю 15.6. В какой ценовой категории вышли подобные ноутбуки?

### 5.3 Что дешевле?

В среднем дешевле ноутбуки с CPU Intel Core i7 7700HQ 2.8GHz или с Intel Core i7 7600U 2.8GHz?



10 минут



## Задание 6.

Найдите самый легкий ноутбук

Но обратите внимание на тип и представление данных в признаке Weight, если что, замените в строке 'kg' на пустую строку через метод `.str.replace()`



5 минут



# Ваши вопросы?

Подведем итоги



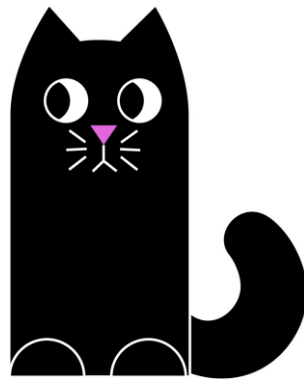


# Домашнее задание



## Домашнее задание 1

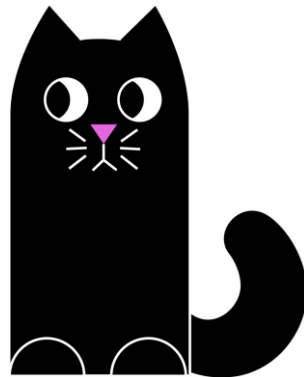
- ☐ Скачать датасет kc-house-data.csv
- ☐ Считать данные с помощью pandas
- ☐ Вывести на экран первые 5 строк
- ☐ Посмотреть на описание признаков и на их содержание



## Домашнее задание 2

Проведите первичный анализ данных

1. Изучите типы данных
2. Найдите количество пропущенных ячеек в данных
3. Посчитайте основные статистики по всем признакам и поизучайте их
4. Пишите выводы



## Домашнее задание 3

Ответьте на несколько вопросов

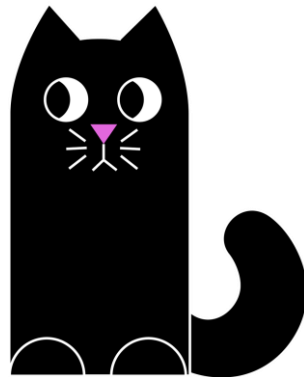
3.1 В каком диапазоне изменяются стоимости недвижимости?

3.2 Какую долю в среднем занимают жилая площадь от всей площади по всем домам?

3.3 Как много домов с разными этажами в данных?

3.4 Насколько хорошие состояния у домов в данных?

3.5 Найдите года, когда построили первый дом, когда построили последний дом в данных?



## Домашнее задание 4

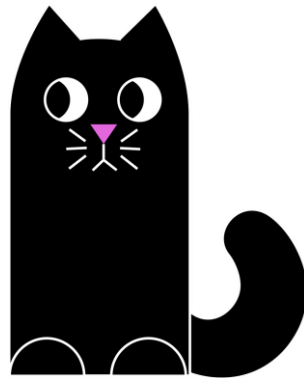
Ответьте на несколько вопросов

4.1 Сколько в среднем стоят дома, у которых 2 спальни?

4.2 Какая в среднем общая площадь домов, у которых стоимость больше 600 000?

4.3 Как много домов коснулся ремонт?

4.4 Насколько в среднем стоимость домов с оценкой grade домов выше 10 отличается от стоимости домов с оценкой grade меньше 4?





## Домашнее задание 5

### 5.1 Выберите дом клиенту

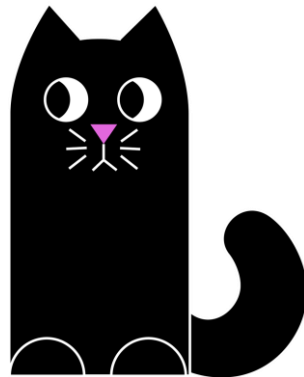
Клиент хочет дом с видом на набережную, как минимум с тремя ванными и с подвалом. Сколько вариантов есть у клиента?

### 5.2 Выберите дом клиенту

Клиент хочет дом либо с очень красивым видом из окна, либо с видом на набережную, в очень хорошем состоянии и год постройки не меньше 1980 года. В какой ценовом диапазоне будут дома?

### 5.3 Выберите дом клиенту

Клиент хочет дом без подвала, с двумя этажами, стоимостью до 150000. Какая оценка по состоянию у таких домов в среднем?





Спасибо  
за внимание

A yellow smiley face is drawn over the text. It has two vertical lines for eyes positioned over the word 'Спасибо' and a wide, upward-curving arc for a mouth positioned over the word 'за внимание'.