

Анализ данных с Python

Лабораторная работа 1 – Pandas DataFrame и выборки данных

Pandas – программная библиотека на языке Python для обработки и анализа данных. Работа pandas с данными строится поверх библиотеки NumPy, являющейся инструментом более низкого уровня. Предоставляет специальные структуры данных и операции для манипулирования числовыми таблицами и временными рядами.

Объект *DataFrame* лучше всего представлять себе в виде обычной таблицы, ведь *DataFrame* является табличной структурой данных. В любой таблице всегда присутствуют строки и столбцы. Столбцами в объекте *DataFrame* выступают объекты *Series*, строки которых являются их непосредственными элементами.

Поясним на примере. Допустим, мы хотим проанализировать данные о зарплате из лабы №1 про Excel:

| Бригада | П.І.Б. | Розряд | Відпрацьовано днів | Зарплата за 1 день | Всього нараховано | Податок | Одержати |
|---------|----------------|--------|--------------------|--------------------|-------------------|---------|----------|
| №1 | Воронюк В.В. | 4 | 24 | 25 | 600 | 78,00 | 522,00 |
| №1 | Денисенко І.В. | 5 | 26 | 30 | 840 | 109,20 | 730,80 |
| №2 | Дуплій О.В. | 6 | 20 | 40 | 800 | 104,00 | 696,00 |
| №2 | Іванов І.І. | 6 | 24 | 40 | 960 | 124,80 | 835,20 |
| №2 | Капітан Г.І. | 4 | 25 | 25 | 650 | 84,50 | 565,50 |
| №1 | Карпова О.П. | 5 | 25 | 30 | 780 | 101,40 | 678,60 |
| №2 | Кірієнко В.Н. | 5 | 26 | 30 | 840 | 109,20 | 730,80 |
| №1 | Коваленко В.Г. | 6 | 24 | 40 | 960 | 124,80 | 835,20 |
| №1 | Луговий А.І. | 4 | 20 | 25 | 500 | 65,00 | 435,00 |
| №2 | Петренко О.В. | 4 | 27 | 25 | 750 | 97,50 | 652,50 |
| №3 | Петров П.П. | 5 | 12 | 30 | 360 | 46,80 | 313,20 |
| №3 | Савчук А.Н. | 5 | 24 | 30 | 720 | 93,60 | 626,40 |
| №2 | Сорокіна Т.П. | 5 | 24 | 30 | 720 | 93,60 | 626,40 |
| №3 | Старчук С.С. | 6 | 23 | 40 | 920 | 119,60 | 800,40 |
| №3 | Шульга О.В. | 6 | 24 | 40 | 960 | 124,80 | 835,20 |

Выбросим некоторые столбцы:

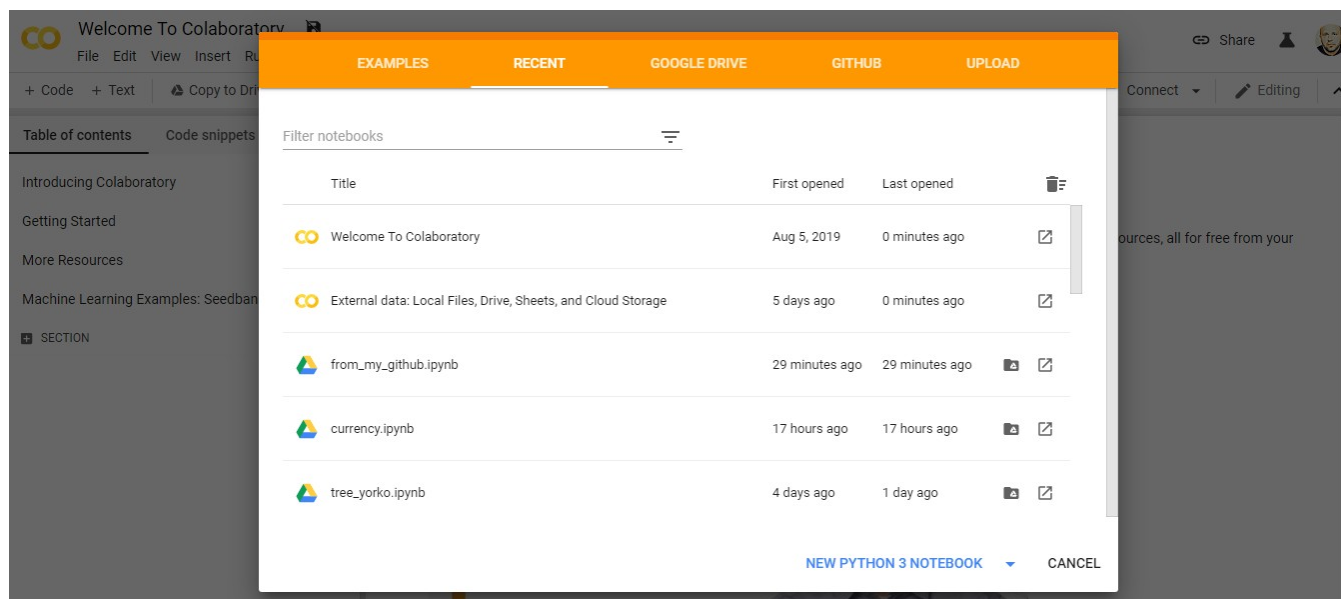
| Бригада | П.І.Б. | Розряд | Відпрацьовано днів | Всього нараховано |
|---------|----------------|--------|--------------------|-------------------|
| №1 | Воронюк В.В. | 4 | 24 | 600 |
| №1 | Денисенко І.В. | 5 | 26 | 840 |
| №2 | Дуплій О.В. | 6 | 20 | 800 |
| №2 | Іванов І.І. | 6 | 24 | 960 |
| №2 | Капітан Г.І. | 4 | 25 | 650 |
| №1 | Карпова О.П. | 5 | 25 | 780 |
| №2 | Кірієнко В.Н. | 5 | 26 | 840 |
| №1 | Коваленко В.Г. | 6 | 24 | 960 |
| №1 | Луговий А.І. | 4 | 20 | 500 |
| №2 | Петренко О.В. | 4 | 27 | 750 |
| №3 | Петров П.П. | 5 | 12 | 360 |
| №3 | Савчук А.Н. | 5 | 24 | 720 |
| №2 | Сорокіна Т.П. | 5 | 24 | 720 |
| №3 | Старчук С.С. | 6 | 23 | 920 |
| №3 | Шульга О.В. | 6 | 24 | 960 |

Таким образом, имеем 5 колонок и 15 строк с данными.

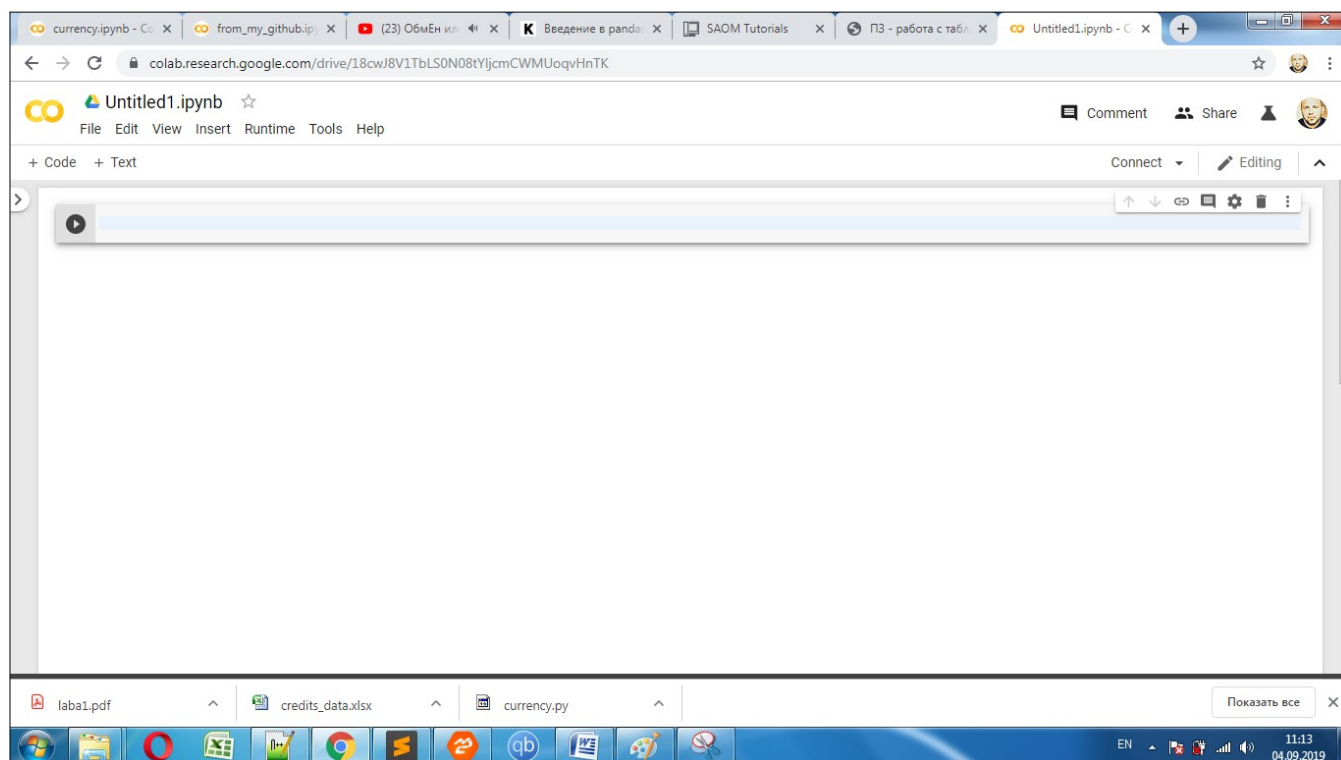
Работать будем в онлайн-сервисе Google Colab (<https://colab.research.google.com>)

Для этого необходимо иметь аккаунт Google, в котором нужно авторизоваться.

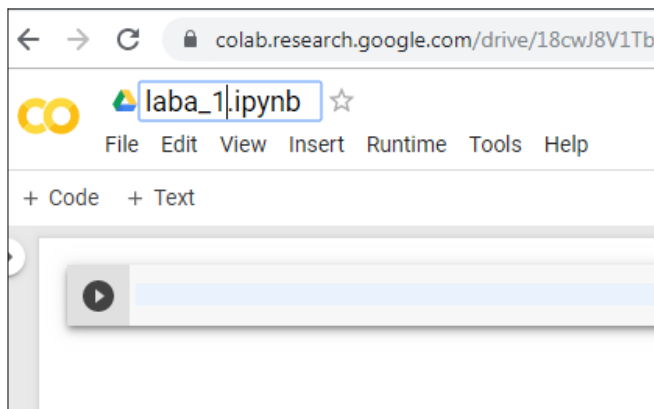
Заходим на сайт, выбираем «NEW PYTHON 3 NOTEBOOK»



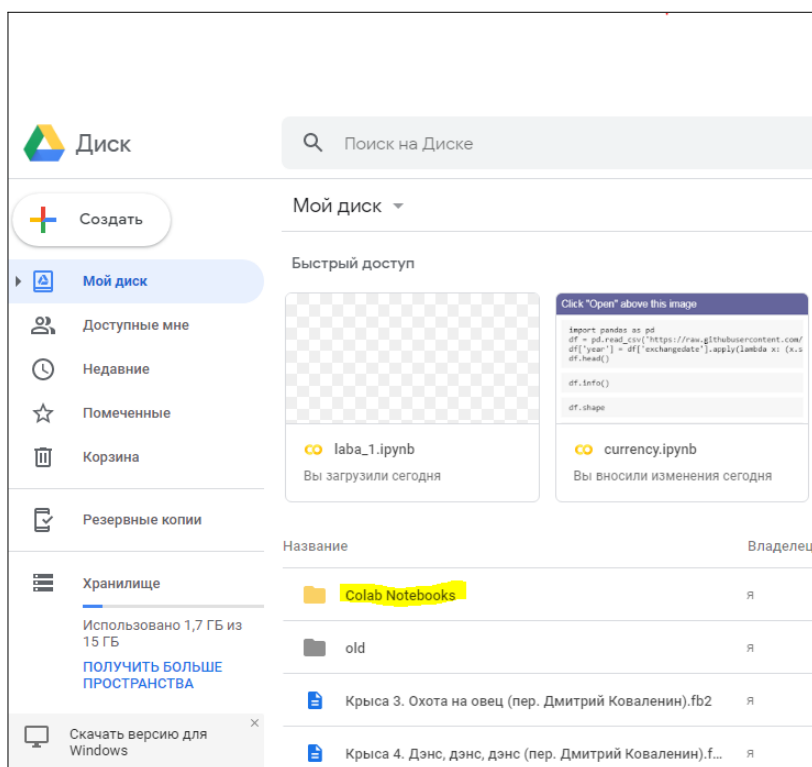
Открывается новый jupyter notebook с именем Untitled1.ipynb.



Переименуем его в «laba_1.ipynb»



Теперь при сохранении файл будет записан на Google-диск в папку «Colab notebooks»



Начинаем кодить. Нам нужно импортировать модули для работы и создать датафрейм на основе нескольких списков (по одному на столбец таблицы).

Будем использовать такие названия полей: «team», «name», «category», «days», «total» для «Бригада», «П.І.Б.», «Розряд», «Відпрацьовано днів», «Всього нараховано» соответственно.

В ячейку листа введем:

```
import pandas as pd

team = [1, 1, 2, 2, 2, 1, 2, 1, 1, 2, 3, 3, 2, 3, 3]

name = ['Воронюк', 'Денисенко', 'Дуплій', 'Іванов', 'Капітан',
        'Карпова', 'Кірієнко', 'Коваленко', 'Луговий', 'Петренко',
        'Петров', 'Савчук', 'Сорокіна', 'Старчук', 'Шульга']

category = [4, 5, 6, 6, 4, 5, 5, 6, 4, 4, 5, 5, 5, 6, 6]

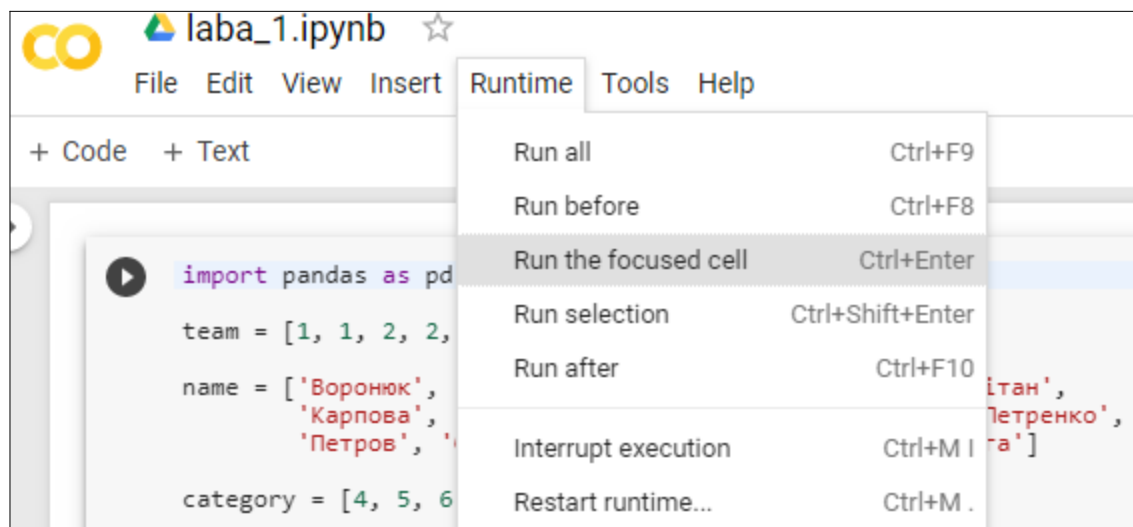
days = [24, 26, 20, 24, 25, 25, 26, 24, 20, 27, 12, 24, 24, 23, 24]

total = [600.0, 840.0, 800.0, 960.0, 650.0, 780.0, 840.0, 960.0, 500.0,
        750.0, 360.0, 720.0, 720.0, 920.0, 960.0]

df = pd.DataFrame( data = {
    'team': team,
    'name': name,
    'category': category,
    'days': days,
    'total': total
})

df
```

Для запуска текущей ячейки нажмем Ctrl+Enter либо «Run the focused cell»



И под ней мы должны увидеть нашу таблицу:

```

    'category': category,
    'days': days,
    'total': total
})
df

```

| | team | name | category | days | total |
|----|------|-----------|----------|------|-------|
| 0 | 1 | Воронюк | 4 | 24 | 600 |
| 1 | 1 | Денисенко | 5 | 26 | 840 |
| 2 | 2 | Дуплій | 6 | 20 | 800 |
| 3 | 2 | Іванов | 6 | 24 | 960 |
| 4 | 2 | Капітан | 4 | 25 | 650 |
| 5 | 1 | Карпова | 5 | 25 | 780 |
| 6 | 2 | Кірієнко | 5 | 26 | 840 |
| 7 | 1 | Коваленко | 6 | 24 | 960 |
| 8 | 1 | Луговий | 4 | 20 | 500 |
| 9 | 2 | Петренко | 4 | 27 | 750 |
| 10 | 3 | Петров | 5 | 12 | 360 |
| 11 | 3 | Савчук | 5 | 24 | 720 |
| 12 | 2 | Сорокіна | 5 | 24 | 720 |
| 13 | 3 | Старчук | 6 | 23 | 920 |
| 14 | 3 | Шульга | 6 | 24 | 960 |

Обратим внимание, что датафрейм проиндексировал строки, начиная с 0.

Теперь создадим несколько новых ячеек для кода (кнопка «+Code»)

| | | |
|--------|--------------|-------|
| CO | laba_1.ipynb | ☆ |
| File | Edit | View |
| Insert | Runtime | Tools |
| Help | | |
| + Code | + Text | |

```

    'category': category,
    'days': days,
    'total': total

```

В следующей ячейке выведем размеры таблицы в виде кортежа двух значений – строки и столбцы – команда **df.shape**

| | | | | | |
|----|---|----------|---|----|-----|
| 12 | 2 | Сорокіна | 5 | 24 | 720 |
| 13 | 3 | Старчук | 6 | 23 | 920 |
| 14 | 3 | Шульга | 6 | 24 | 960 |

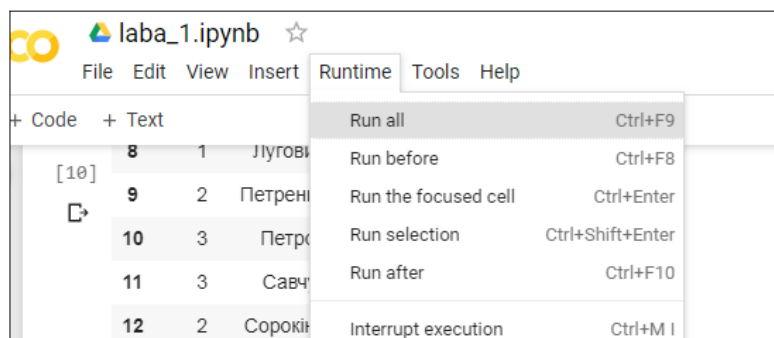
```

df.shape

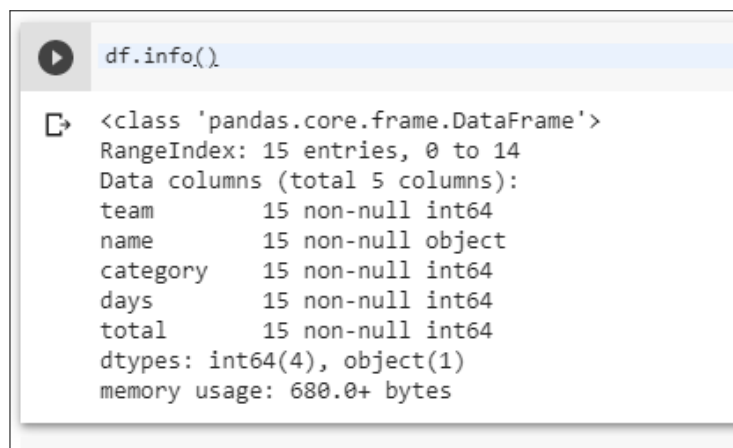
```

(15, 5)

Как запустить код в текущей ячейке, мы знаем, а чтобы перезапустить все ячейки с начала документа, используем Ctrl+F9 или соответствующий пункт меню



Идем дальше. Выведем инфу о датафрейме – команда `df.info()`

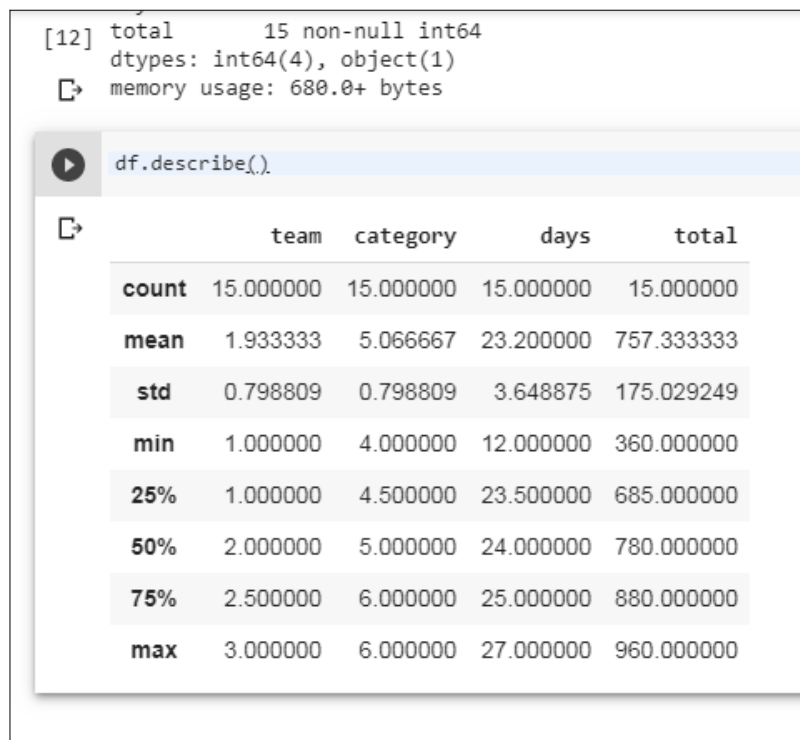


Здесь мы можем видеть типы данных pandas в колонках:

целочисленные (int64) – team, category, days, total

строковые (object) – name

Команда `df.describe()` выведет основные характеристики для числовых данных



count – количество значений

mean – среднее значение

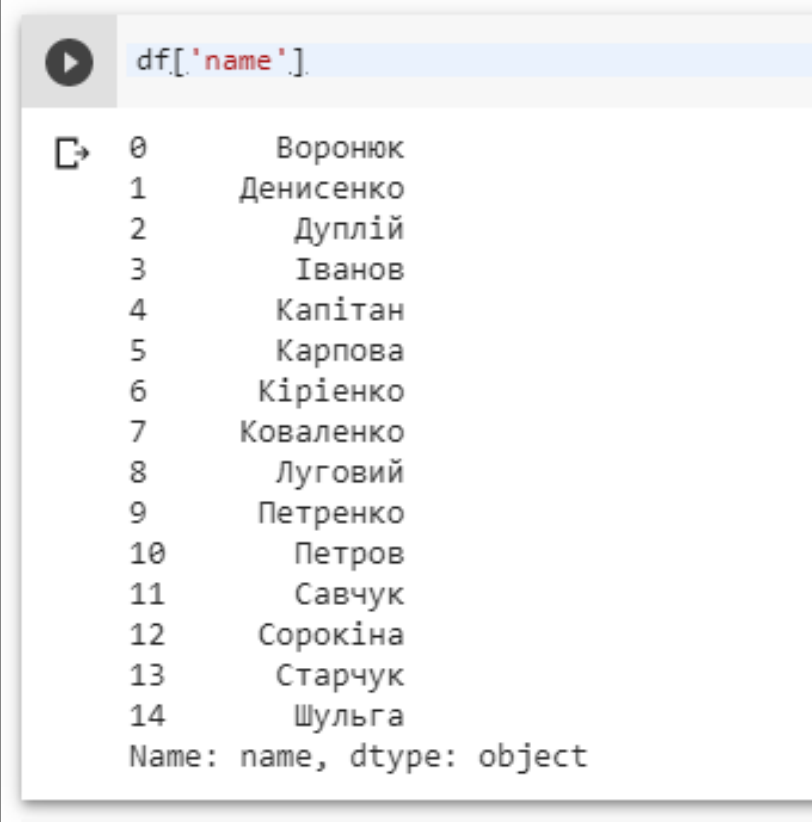
std – среднее отклонение от среднего значения

min, max – понятно

25%, 50%, 75% – значения, близкие к 1-й, 2-й и 3-й четвертям диапазона между минимумом и максимумом

Перейдем к выборке данных. Если мы хотим отобразить только один столбец с фамилиями, то:

```
df['name']
```



The screenshot shows a Jupyter Notebook interface. At the top, there is a code cell with the text `df['name']`. Below the code cell, the output is displayed as a list of 15 names, indexed from 0 to 14. The names are: Воронюк, Денисенко, Дуплій, Іванов, Капітан, Карпова, Кірієнко, Коваленко, Луговий, Петренко, Петров, Савчук, Сорокіна, Старчук, and Шульга. At the bottom of the output, it says "Name: name, dtype: object".

| 0 | Воронюк |
|----|-----------|
| 1 | Денисенко |
| 2 | Дуплій |
| 3 | Іванов |
| 4 | Капітан |
| 5 | Карпова |
| 6 | Кірієнко |
| 7 | Коваленко |
| 8 | Луговий |
| 9 | Петренко |
| 10 | Петров |
| 11 | Савчук |
| 12 | Сорокіна |
| 13 | Старчук |
| 14 | Шульга |

Name: name, dtype: object

Если несколько полей, например фамилии и зарплата:

```
df[['name', 'total']]
```



The screenshot shows a Jupyter Notebook interface. At the top, a code cell contains the command `df[['name', 'total']]`. Below the code cell, the output is displayed as a table with two columns: 'name' and 'total'. The table contains 15 rows of data, indexed from 0 to 14. The names are in Ukrainian, and the total values are numerical.

| | name | total |
|----|-----------|-------|
| 0 | Воронюк | 600 |
| 1 | Денисенко | 840 |
| 2 | Дуплій | 800 |
| 3 | Іванов | 960 |
| 4 | Капітан | 650 |
| 5 | Карпова | 780 |
| 6 | Кірієнко | 840 |
| 7 | Коваленко | 960 |
| 8 | Луговий | 500 |
| 9 | Петренко | 750 |
| 10 | Петров | 360 |
| 11 | Савчук | 720 |
| 12 | Сорокіна | 720 |
| 13 | Старчук | 920 |
| 14 | Шульга | 960 |

Обратите внимание, что в этом случае мы используем двойные квадратные скобки (передаем список названий полей).

Для трех полей было бы, например так:

```
df[['name', 'category', 'total']]
```

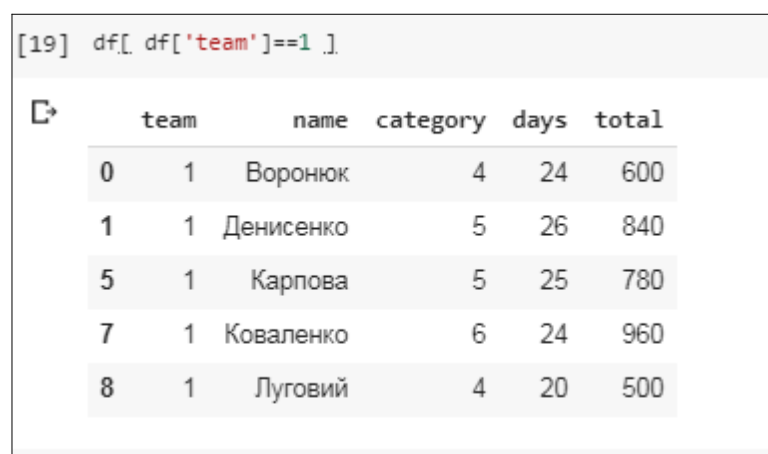


```
df[['name', 'category', 'total']]
```

| | name | category | total |
|----|-----------|----------|-------|
| 0 | Воронюк | 4 | 600 |
| 1 | Денисенко | 5 | 840 |
| 2 | Дуплій | 6 | 800 |
| 3 | Іванов | 6 | 960 |
| 4 | Капітан | 4 | 650 |
| 5 | Карпова | 5 | 780 |
| 6 | Кірієнко | 5 | 840 |
| 7 | Коваленко | 6 | 960 |
| 8 | Луговий | 4 | 500 |
| 9 | Петренко | 4 | 750 |
| 10 | Петров | 5 | 360 |
| 11 | Савчук | 5 | 720 |
| 12 | Сорокіна | 5 | 720 |
| 13 | Старчук | 6 | 920 |
| 14 | Шульга | 6 | 960 |

Попробуем сделать выборку по строкам, например только первую бригаду:

```
df[ df['team']==1 ]
```

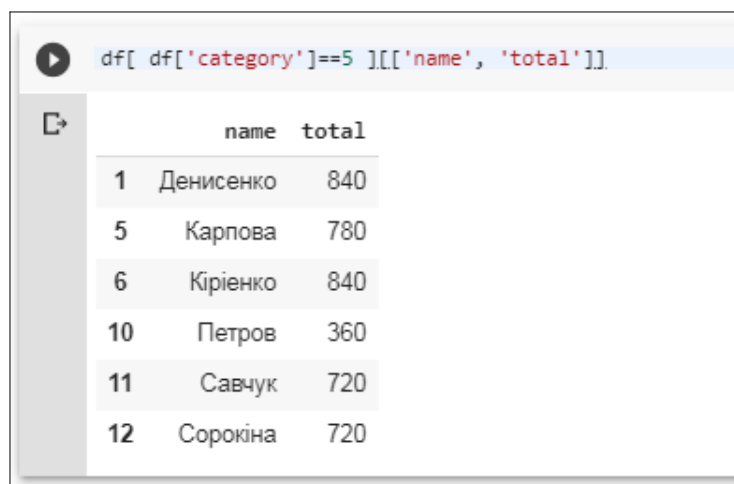


```
[19] df[ df['team']==1 ]
```

| | team | name | category | days | total |
|---|------|-----------|----------|------|-------|
| 0 | 1 | Воронюк | 4 | 24 | 600 |
| 1 | 1 | Денисенко | 5 | 26 | 840 |
| 5 | 1 | Карпова | 5 | 25 | 780 |
| 7 | 1 | Коваленко | 6 | 24 | 960 |
| 8 | 1 | Луговий | 4 | 20 | 500 |

Ну, и теперь выборка по двум направлениям (пятый разряд, фамилии и зарплата):

```
df[ df['category']==5 ][['name', 'total']]
```



```
df[ df['category']==5 ][['name', 'total']]
```

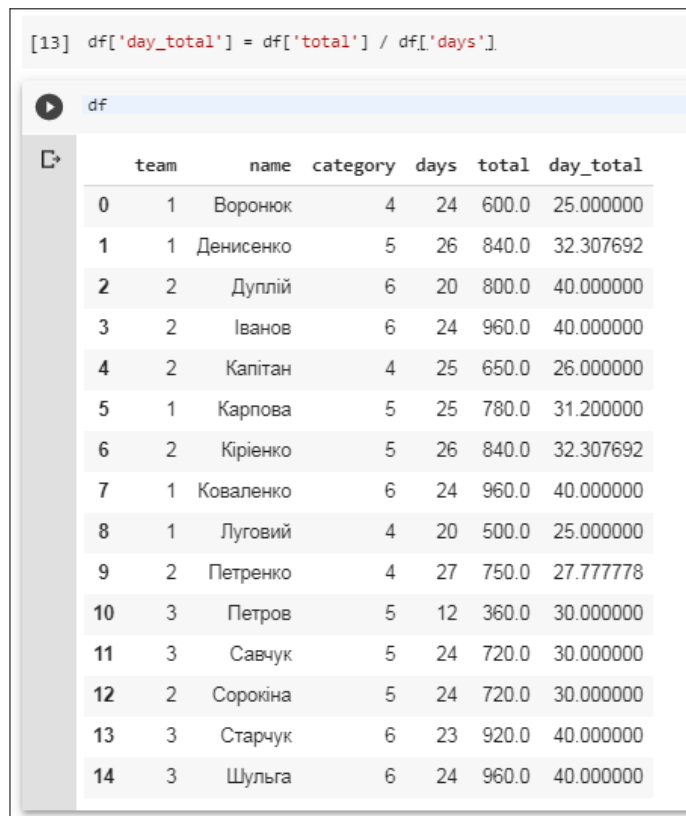
| | name | total |
|----|-----------|-------|
| 1 | Денисенко | 840 |
| 5 | Карпова | 780 |
| 6 | Кірієнко | 840 |
| 10 | Петров | 360 |
| 11 | Савчук | 720 |
| 12 | Сорокіна | 720 |

И последнее в этой работе: создадим вычисляемое поле и добавим его в датафрейм. Допустим, мы хотим в отдельную колонку вынести зарплату за 1 отработанный день. Очевидно, что формула вычисления будет: суммарная зарплата делить на количество дней.

Создадим поле 'day_total':

```
df['day_total'] = df['total'] / df['days']
```

и выведем датафрейм на экран:



```
[13] df['day_total'] = df['total'] / df['days']
```

```
df
```

| | team | name | category | days | total | day_total |
|----|------|-----------|----------|------|-------|-----------|
| 0 | 1 | Воронюк | 4 | 24 | 600.0 | 25.000000 |
| 1 | 1 | Денисенко | 5 | 26 | 840.0 | 32.307692 |
| 2 | 2 | Дуплій | 6 | 20 | 800.0 | 40.000000 |
| 3 | 2 | Іванов | 6 | 24 | 960.0 | 40.000000 |
| 4 | 2 | Капітан | 4 | 25 | 650.0 | 26.000000 |
| 5 | 1 | Карпова | 5 | 25 | 780.0 | 31.200000 |
| 6 | 2 | Кірієнко | 5 | 26 | 840.0 | 32.307692 |
| 7 | 1 | Коваленко | 6 | 24 | 960.0 | 40.000000 |
| 8 | 1 | Луговий | 4 | 20 | 500.0 | 25.000000 |
| 9 | 2 | Петренко | 4 | 27 | 750.0 | 27.777778 |
| 10 | 3 | Петров | 5 | 12 | 360.0 | 30.000000 |
| 11 | 3 | Савчук | 5 | 24 | 720.0 | 30.000000 |
| 12 | 2 | Сорокіна | 5 | 24 | 720.0 | 30.000000 |
| 13 | 3 | Старчук | 6 | 23 | 920.0 | 40.000000 |
| 14 | 3 | Шульга | 6 | 24 | 960.0 | 40.000000 |

Индивидуальное задание

Необходимо создать датафрейм, взяв за основу данные из 1-й работы по Excel, вывести информацию о нем, сделать выборки данных и добавить 2 вычисляемых поля. Код сопроводить комментариями.