

## ▼ Pandas

Материалы:

- Макрушин С.В. "Лекция 2: Библиотека Pandas"
- [https://pandas.pydata.org/docs/user\\_guide/index.html#](https://pandas.pydata.org/docs/user_guide/index.html#)
- <https://pandas.pydata.org/docs/reference/index.html>
- Уэс Маккини. Python и анализ данных

```
import pandas as pd
import numpy as np
from google.colab import drive
drive.mount('/content/gdrive')
```

Drive already mounted at /content/gdrive; to attempt to forcibly remount, call drive.mou



## ► Задачи для совместного разбора

[ ] ↳ 11 cells hidden

## ▼ Лабораторная работа №2

### ▼ Базовые операции с DataFrame

1.1 В файлах `recipes_sample.csv` и `reviews_sample.csv` находится информация об рецептах блюд и отзывах на эти рецепты соответственно. Загрузите данные из файлов в виде `pd.DataFrame` с названиями `recipes` и `reviews`. Обратите внимание на корректное считывание столбца с индексами в таблице `reviews` (безымянный столбец).

```
recipes0 = pd.read_csv('gdrive/My Drive/Colab Notebooks/data/recipes_sample.csv', header = 0)
recipes = pd.DataFrame(recipes0)
recipes.head()
```

	name	id	minutes	contributor_id	submitted	n_steps	description	n_ingredient
0	george s at the cove black bean	44123	90	35193	2002-10-25	NaN	an original recipe created by chef scott meska...	1

```
reviews0 = pd.read_csv('gdrive/My Drive/Colab Notebooks/data/reviews_sample.csv', header = 0)
reviews = pd.DataFrame(reviews0)
reviews.head()
```

	Unnamed: 0	user_id	recipe_id	date	rating	
0	370476	21752	57993	2003-05-01	5	Last week whole sides of frozen
1	624300	431813	142201	2007-09-16	5	So simple and so tasty! I used a
2	187037	400708	252013	2008-01-10	4	Very nice breakfast HH, easy to make
3	706134	2001852463	404716	2017-12-11	5	These are a favorite for the holid
4	312179	95810	129396	2008-03-14	5	Excellent soup! The tomato flav

1.2 Для каждой из таблиц выведите основные параметры:

- количество точек данных (строк);
- количество столбцов;
- тип данных каждого столбца.

```
recipes.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 8 columns):
#   Column                Non-Null Count  Dtype
---  -
0   name                  30000 non-null object
1   id                    30000 non-null int64
2   minutes               30000 non-null int64
3   contributor_id        30000 non-null int64
4   submitted             30000 non-null object
5   n_steps               18810 non-null float64
6   description           29377 non-null object
7   n_ingredients         21120 non-null float64
dtypes: float64(2), int64(3), object(3)
memory usage: 1.8+ MB
```

```
reviews.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 126696 entries, 0 to 126695
Data columns (total 6 columns):
#   Column      Non-Null Count  Dtype
---  -
0   Unnamed: 0   126696 non-null  int64
1   user_id      126696 non-null  int64
2   recipe_id    126696 non-null  int64
3   date         126696 non-null  object
4   rating       126696 non-null  int64
5   review       126679 non-null  object
dtypes: int64(4), object(2)
memory usage: 5.8+ MB
```

1.3 Исследуйте, в каких столбцах таблиц содержатся пропуски. Посчитайте долю строк, содержащих пропуски, в отношении к общему количеству строк.

```
print(reviews.isnull().sum())
```

```
Unnamed: 0      0
user_id         0
recipe_id       0
date            0
rating          0
review         17
dtype: int64
```

```
print(recipes.isnull().sum())
```

```
name           0
id             0
minutes        0
contributor_id 0
submitted      0
n_steps       11190
description     623
n_ingredients  8880
dtype: int64
```

```
print('reviews: ', '%.3f' % (reviews.isnull().sum().sum()/reviews.shape[0]*100), '%')
print('recipes: ', '%.3f' % (recipes.isnull().sum().sum()/recipes.shape[0]*100), '%')
```

```
reviews:  0.013 %
recipes:  68.977 %
```

1.4 Рассчитайте среднее значение для каждого из числовых столбцов (где это имеет смысл).

```
reviews[['Unnamed: 0', 'rating']].mean(numeric_only=None)
```

```
Unnamed: 0    566089.821147
rating         4.410802
dtype: float64
```

```
recipes[['minutes', 'n_steps', 'n_ingredients']].mean(numeric_only=True)
```

```
minutes         123.358133
n_steps          9.805582
n_ingredients    9.008286
dtype: float64
```

### 1.5 Создайте серию из 10 случайных названий рецептов.

```
recipes['name'].sample(n=10)
```

```
20862    pineapple cabbage salad
3267      bleu cheese bread
7852      crabmeat puffs
1730    baked acorn squash with spicy maple syrup
22204      raspberry sorbet
5377      cherry amaretto dessert low carb
20720      pide  turkish flat cake
5132      cheese filled peppers
7754      country style scrambled eggs
24171      silky lemon pudding
Name: name, dtype: object
```

### 1.6 Измените индекс в таблице reviews, пронумеровав строки, начиная с нуля.

```
reviews = reviews.rename(columns={"Unnamed: 0": "index"})
```

```
reviews['index'] = pd.Series(np.arange(reviews.shape[0]))
reviews.head()
```

	index	user_id	recipe_id	date	rating	review
0	0	21752	57993	2003-05-01	5	Last week whole sides of frozen salmon fillet ...
1	1	431813	142201	2007-09-16	5	So simple and so tasty! I used a yellow capsic...
2	2	400708	252013	2008-01-10	4	Very nice breakfast HH, easy to make and yummy...
3	3	333155400	101710	2017-12-	5	These are a favorite for the holidays and so

1.7 Выведите информацию о рецептах, время выполнения которых не больше 20 минут и кол-во ингредиентов в которых не больше 5.

```
recipes.loc[(recipes['minutes']<=20) & (recipes['n_ingredients']<=5)]
```

	name	id	minutes	contributor_id	submitted	n_steps	description	n_ing
28	quick biscuit bread	302399	20	213909	2008-05-06	11.0	this is a wonderful quick bread to make as an ...	
60	peas fit for a king or queen	303944	20	213909	2008-05-16	NaN	this recipe is so simple and the flavors are s...	
90	hawaiian sunrise mimosa	100837	5	58104	2004-09-29	4.0	pineapple mimosa was changed to hawaiian sunri...	
91	tasty dish s banana pudding in 2 minutes	286484	2	47892	2008-02-13	NaN	"mmmm, i love bananas!" a - -tasty dish-- origi...	
94	1 minute meatballs	11361	13	4470	2001-09-03	NaN	this is a real short cut for cooks in a hurry....	
...	...	...	...	...	...	...	...	...
	zip and steam red						i haven't tried	

## ▼ Работа с датами в pandas

2.1 Преобразуйте столбец `submitted` из таблицы `recipes` в формат времени.

Модифицируйте решение задачи 1.1 так, чтобы считать столбец сразу в нужном формате.

```
recipes['submitted'] = pd.to_datetime(recipes['submitted'])
recipes1 = pd.read_csv('gdrive/My Drive/Colab Notebooks/data/recipes_sample.csv', header = 0,
```

```
recipes1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 8 columns):
#   Column                Non-Null Count  Dtype  
---  -
0   name                   30000 non-null  object 
1   id                     30000 non-null  int64  
2   minutes                30000 non-null  int64  
3   contributor_id         30000 non-null  int64  
4   submitted              30000 non-null  datetime64[ns]
5   n_steps                18810 non-null  float64 
6   description            29377 non-null  object 
7   n_ingredients          21120 non-null  float64 
dtypes: datetime64[ns](1), float64(2), int64(3), object(2)
memory usage: 1.8+ MB
```

2.2 Выведите информацию о рецептах, добавленных в датасет не позже 2010 года.

```
recipes.loc[(recipes['submitted']>='2010-01-01')]
```

name id minutes contributor id submitted n steps

## ▼ Работа со строковыми данными в pandas

3.1 Добавьте в таблицу `recipes` столбец `description_length`, в котором хранится длина описания рецепта из столбца `description`.

```

# desc = [len(str(i)) for i in pd.Series(recipes['description']).array]
# pd.Series(recipes['description']).array
# print(desc)
recipes['description_length'] = [len(str(i)) for i in pd.Series(recipes['description']).array]

4/1          cheese  463219      25      383346      22      8.0

recipes.head()

```

	name	id	minutes	contributor_id	submitted	n_steps	description	n_ingredient
0	george s at the cove black bean soup	44123	90	35193	2002-10-25	NaN	an original recipe created by chef scott meska...	1
1	healthy for them yogurt popsicles	67664	10	91970	2003-07-26	NaN	my children and their friends ask for my homem...	1
2	i can t believe it s	38798	30	1533	2002-08-29	NaN	these were so go, it surprised	

3.2 Измените название каждого рецепта в таблице `recipes` таким образом, чтобы каждое слово в названии начиналось с прописной буквы.

```

recipes['name'] = recipes['name'].str.title()

recipes.head()

```

	name	id	minutes	contributor_id	submitted	n_steps	description
0	George S At The Cove Black Bean Soup	44123	90	35193	2002-10-25	NaN	an original recipe created by chef scott meska...
1	Healthy For Them Yogurt Popsicles	67664	10	91970	2003-07-26	NaN	my children and their friends ask for my homem...
2	I Can T Believe It S Spinach	38798	30	1533	2002-08-29	NaN	these were so go, it surprised even me.

3.3 Добавьте в таблицу `recipes` столбец `name_word_count`, в котором хранится количество слов из названии рецепта (считайте, что слова в названии разделяются только пробелами). Обратите внимание, что между словами может располагаться несколько пробелов подряд.

```
4 Air Reef Fondue 84797 25 4470 2004-02- 4 0 very romantic casual
recipes['name_word_count'] = recipes['name'].str.split().str.len()
```

```
recipes.head()
```

	name	id	minutes	contributor_id	submitted	n_steps	description	n_ingredient
0	George S At The Cove Black Bean Soup	44123	90	35193	2002-10-25	NaN	an original recipe created by chef scott meska...	1
1	Healthy For Them Yogurt Popsicles	67664	10	91970	2003-07-26	NaN	my children and their friends ask for my homem...	1
2	I Can T Believe It S Spinach	38798	30	1533	2002-08-29	NaN	these were so go, it surprised even me.	

## ▼ Группировки таблиц `pd.DataFrame`

4.1 Посчитайте количество рецептов, представленных каждым из участников (`contributor_id`). Какой участник добавил максимальное кол-во рецептов?

```
recipes.groupby('contributor_id')['name'].count().sort_values()
```



```

contributor_id
302190      1
449247      1
449229      1
449014      1
448748      1
...
169430     183
1533       186
37779      345
37449      346
89831      421
Name: name, Length: 8404, dtype: int64

```

4.2 Посчитайте средний рейтинг к каждому из рецептов. Для скольких рецептов отсутствуют отзывы? Обратите внимание, что отзыв с нулевым рейтингом или не заполненным текстовым описанием не считается отсутствующим.

```

reviews.groupby('recipe_id')['rating'].mean()

recipe_id
48      1.000000
55      4.750000
66      4.944444
91      4.750000
94      5.000000
...
536547   5.000000
536610   0.000000
536728   4.000000
536729   4.750000
536747   0.000000
Name: rating, Length: 28100, dtype: float64

```

```

len(recipes.groupby('id')['id'].count())-len(reviews.groupby('recipe_id')['recipe_id'].count(
1900

```

4.3 Посчитайте количество рецептов с разбивкой по годам создания.

```

recipes['year'] = recipes['submitted'].dt.year
recipes

```

	name	id	minutes	contributor_id	submitted	n_steps	description
0	George S At The Cove Black Bean Soup	44123	90	35193	2002-10-25	NaN	an original recipe created by chef sco meska.
1	Healthy For Them Yogurt Popsicles	67664	10	91970	2003-07-26	NaN	my children and the friends as for me homem.
2	I Can T Believe It S Spinach	38798	30	1533	2002-08-29	NaN	these were so good, surprise even me
3	Italian Gut Busters	35173	45	22724	2002-07-27	NaN	my sister-in-law made these for us at a family.
4	Love Is In The Air Beef Fondue Sauces	84797	25	4470	2004-02-23	4.0	i think fondue is very romantic casual din.
...	...	...	...	...	...	...	.
29995	Zurie S Holey Rustic Olive And Cheddar Bread	267661	80	200862	2007-11-25	16.0	this is based on a french recipe but changed.
29996	Zwetschkuchen Bavarian Plum Cake	386977	240	177443	2009-08-24	NaN	this is traditional fresh plum cake thought.
29997	Zwiebelkuchen Southwest German Onion Cake	103312	75	161745	2004-11-03	NaN	this is traditional late summer early fall season.
							this is delicious

```
recipes.groupby('year')['id'].count()
```

year	
1999	275
2000	104
2001	589
2002	2644
2003	2334

```
2004    2153
2005    3130
2006    3473
2007    4429
2008    4029
2009    2963
2010    1538
2011     922
2012     659
2013     490
2014     139
2015      42
2016      24
2017      39
2018      24
Name: id, dtype: int64
```

## ▼ Объединение таблиц pd.DataFrame

5.1 При помощи объединения таблиц, создайте DataFrame, состоящий из четырех столбцов: `id`, `name`, `user_id`, `rating`. Рецепты, на которые не оставлен ни один отзыв, должны отсутствовать в полученной таблице. Подтвердите правильность работы вашего кода, выбрав рецепт, не имеющий отзывов, и попытавшись найти строку, соответствующую этому рецепту, в полученном DataFrame.

```
reviews = reviews.rename(columns={"recipe_id":"id"})
```

```
reviews.shape
```

```
(126696, 6)
```

```
df = pd.merge(recipes[['id','name']], reviews[['user_id','rating','id']], on='id', how='inner')
df
```

	id	name	user_id	rating	
0	44123	George S At The Cove Black Bean Soup	743566	5	
1	44123	George S At The Cove Black Bean Soup	76503	5	
2	44123	George S At The Cove Black Bean Soup	34206	5	
3	67664	Healthy For Them Yogurt Popsicles	494084	5	
4	67664	Healthy For Them Yogurt Popsicles	666445	5	

```
new_df = pd.merge(recipes[['id','name']], reviews[['user_id','rating','id']], on='id', how='left')
new_df[new_df['rating'].isna()]
```

	id	name	user_id	rating	
191	223349	Pasta With Shrimp Eggplant	NaN	NaN	
195	342620	Secret Ingredient Bbq Meatloaf	NaN	NaN	
216	276594	Windy S Sweet And Sour Meatballs	NaN	NaN	
282	216068	Goulashy Beef Stew For The Slow Cooker	NaN	NaN	
300	306590	Old Bay Grilled Steak Fries	NaN	NaN	
...	...	...	...	...	
128149	66966	Zingy Potato Salad	NaN	NaN	
128256	249846	Zucchini Mock Green Papaya Salad	NaN	NaN	
128273	256177	Zucchini And Corn With Cheese	NaN	NaN	
128436	178243	Zucchini Nut Salad	NaN	NaN	
128535	363362	Zucchini With Bell Pepper And Tomato	NaN	NaN	

1900 rows × 4 columns

```
df[df.id == '223349']
```

id	name	user_id	rating	
----	------	---------	--------	--

5.2 При помощи объединения таблиц и группировок, создайте DataFrame, состоящий из трех столбцов: `recipe_id`, `name`, `review_count`, где столбец `review_count` содержит кол-во отзывов, оставленных на рецепт `recipe_id`. У рецептов, на которые не оставлен ни один отзыв, в столбце `review_count` должен быть указан 0. Подтвердите правильность работы вашего кода, выбрав рецепт, не имеющий отзывов, и найдя строку, соответствующую этому рецепту, в полученном DataFrame.

```

review_counter = pd.merge(recipes[['id','name']], reviews[['rating','id']], on='id', how='left')
review_counter = review_counter.rename(columns={"id":"recipe_id"})
review_new = review_counter.groupby(['recipe_id','name']).agg(review_count=('rating','count'))
review_new

```

	recipe_id	name	review_count	
0	48	Boston Cream Pie	2	
1	55	Betty Crocker S Southwestern Guacamole Dip	4	
2	66	Black Coffee Barbecue Sauce	18	
3	91	Brown Rice And Vegetable Pilaf	4	
4	94	Blueberry Buttermilk	4	
...	...	...	...	
29995	536547	Cauliflower Ceviche	1	
29996	536610	Miracle Home Made Puff Pastry	1	
29997	536728	Gluten Free Vegemite	1	
29998	536729	Creole Watermelon Feta Salad	4	
29999	536747	Lemon Pom Pom Cake	4	

30000 rows × 3 columns

### 5.3. Выясните, рецепты, добавленные в каком году, имеют наименьший средний рейтинг?

```
pd.merge(df, recipes[['id', 'year']], on='id', how='left').groupby(['year'])['rating'].mean()
```

```

year
2017    2.750000
2018    3.388889
2016    3.538462
2015    4.207317
1999    4.274895
2000    4.284585
2013    4.336508
2009    4.360447
2011    4.375850
2008    4.387416
2001    4.393945
2002    4.404645
2010    4.406279
2005    4.409687
2012    4.412736
2006    4.416310
2007    4.420073
2003    4.439152
2004    4.463778

```

```
2014    4.538235
```

```
Name: rating, dtype: float64
```

## ▼ Сохранение таблиц `pd.DataFrame`

6.1 Отсортируйте таблицу в порядке убывания величины столбца `name_word_count` и сохраните результаты выполнения заданий 3.1-3.3 в csv файл.

```
recipes = recipes.sort_values(by='name_word_count', ascending = False)
recipes
```

	name	id	minutes	contributor_id	submitted	n_steps	descri
26223	Subru Uncle S Whole Green Moong Dal I LI Be Ma...	77188	95	6357	2003-11-21	NaN	my dad and morn enjoy this lentil cu
28083	Tsr Version Of T G I Friday S Black Bean	102274	75	74652	2004-10-19	9.0	www.topsecretrecipe i got this cc

```
recipes.to_csv('gdrive/My Drive/Colab Notebooks/data/recipes_new.csv')
```

Subru Uncle  
S Whole  
Green  
Moong Dal I  
LI Be Ma...

6.2 Воспользовавшись `pd.ExcelWriter`, сохраните результаты 5.1 и 5.2 в файл: на лист с названием `Рецепты с оценками` сохраните результаты выполнения 5.1; на лист с названием `Количество отзывов по рецептам` сохраните результаты выполнения 5.2.

```
27876 Version Of I 113346 20 175727 2003-03- 5.0 this recipe is top  
with pd.ExcelWriter("gdrive/My Drive/Colab Notebooks/data/df_new.xlsx") as writer:  
df.to_excel(writer, sheet_name="Рецепты с оценками")  
review_new.to_excel(writer, sheet_name="Количество отзывов по рецептам")
```

5734	Curry Or Cat S Vomit On A Bed Of Magg...	294898	30	802799	2008-03-28	11.0	an old family recipe easy to make s
...	...	...	...	...	...	...	...
3253	Blackmoons	323195	430	415934	2008-09-04	5.0	my mom was a new in the 1950s whe
4138	Bushwhacker	156521	10	177392	2006-02-17	1.0	this drink is an exc after dinner d
2357	Basbousa	12957	60	18391	2001-10-20	NaN	this is a traditional r eastern dess
15052	Josefinas	264859	20	498271	2007-11-11	7.0	from the junior lea corpus christi