Pandas

Материалы:

- Макрушин С.В. "Лекция 2: Библиотека Pandas"
- https://pandas.pydata.org/docs/user_guide/index.html#
- https://pandas.pydata.org/docs/reference/index.html
- Уэс Маккини. Python и анализ данных

In [4]:
```
import pandas as pd
import numpy as np
```

Задачи для совместного разбора

1. Загрузите данные из файла `sp500hst.txt` и обозначьте столбцы в соответствии с содержимым: `"date", "ticker", "open", "high", "low", "close", "volume"`.

1. Рассчитайте среднее значение показателей для каждого из столбцов с номерами 3-6.

1. Добавьте столбец, содержащий только число месяца, к которому относится дата.

1. Рассчитайте суммарный объем торгов для для одинаковых значений тикеров.

1. Загрузите данные из файла sp500hst.txt и обозначьте столбцы в соответствии с содержимым: "date", "ticker", "open", "high", "low", "close", "volume". Добавьте столбец с расшифровкой названия тикера, используя данные из файла `sp_data2.csv`. В случае нехватки данных об именах тикеров корректно обработать их.

Лабораторная работа №2.1

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).

In [168…
```
data_recipes = pd.read_csv("recipes_sample.csv", parse_dates=["submitted"])
data_recipes.head()
```

Out[168]:

	name	id	minutes	contributor_id	submitted	n_steps	description	n_ingredients
0	george s at the cove black bean soup	44123	90	35193	2002-10-25	NaN	an original recipe created by chef scott meska...	18.0
1	healthy for them yogurt popsicles	67664	10	91970	2003-07-26	NaN	my children and their friends ask for my homem...	NaN
2	i can t believe it s spinach	38798	30	1533	2002-08-29	NaN	these were so go, it surprised even me.	8.0
3	italian gut busters	35173	45	22724	2002-07-27	NaN	my sister-in-law made these for us at a family...	NaN
4	love is in the air beef fondue sauces	84797	25	4470	2004-02-23	4.0	i think a fondue is a very romantic casual din...	NaN

In [170…
```
data_reviews = pd.read_csv("reviews_sample.csv", index_col=0)
data_reviews.head()
```

Out[170]:

	user_id	recipe_id	date	rating	review
370476	21752	57993	2003-05-01	5	Last week whole sides of frozen salmon fillet ...
624300	431813	142201	2007-09-16	5	So simple and so tasty! I used a yellow capsi...
187037	400708	252013	2008-01-10	4	Very nice breakfast HH, easy to make and yummy...
706134	2001852463	404716	2017-12-11	5	These are a favorite for the holidays and so e...
312179	95810	129396	2008-03-14	5	Excellent soup! The tomato flavor is just gre...

1.2 Для каждой из таблиц выведите основные параметры:

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

In [173…
```
data_reviews.shape[0], data_recipes.shape[0]
```

Out[173]:
```
(126696, 30000)
```

In [175…
```
data_reviews.shape[1], data_recipes.shape[1]
```

Out[175]:
```
(5, 8)
```

In [177…
```
data_reviews.dtypes
```

Out[177]:
```
user_id       int64
recipe_id     int64
date         object
rating        int64
review       object
dtype: object
```

In [179…
```
data_recipes.dtypes
```

Out[179]:
```
name                      object
id                         int64
minutes                    int64
contributor_id             int64
submitted         datetime64[ns]
n_steps                  float64
description               object
n_ingredients            float64
dtype: object
```

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

In [182…
```
missing_values_recipes = data_recipes.isnull().sum()
columns_with_missing_recipes = missing_values_recipes[missing_values_recipes > 0]


missing_rows_ratio = data_recipes.isnull().any(axis=1).mean()

print("Столбцы с пропущенными значениями и их количество:")
print(columns_with_missing_recipes)
print(f"\nДоля строк с пропусками: {missing_rows_ratio:.2%}")
```

```
Столбцы с пропущенными значениями и их количество:
n_steps          11190
description        623
n_ingredients     8880
dtype: int64

Доля строк с пропусками: 56.85%
```

In [184…
```
missing_values_reviews = data_reviews.isnull().sum()
columns_with_missing_reviews = missing_values_reviews[missing_values_reviews > 0]


missing_rows_ratio = data_reviews.isnull().any(axis=1).mean()

print("Столбцы с пропущенными значениями и их количество:")
print(columns_with_missing_reviews)
print(f"\nДоля строк с пропусками: {missing_rows_ratio:.2%}")
```

```
Столбцы с пропущенными значениями и их количество:
review    17
dtype: int64

Доля строк с пропусками: 0.01%
```

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

In [187…
```
mean_values_recipes = data_recipes.mean(numeric_only=True)
```

In [187…
```
mean_values_recipes = data_recipes.mean(numeric_only=True)
mean_values_reviews = data_reviews.mean(numeric_only=True)

mean_values_recipes
```

Out[187]:
```
id               2.218793e+05
minutes          1.233581e+02
contributor_id   5.635901e+06
n_steps          9.805582e+00
n_ingredients    9.008286e+00
dtype: float64
```

In [189…
```
mean_values_reviews
```

Out[189]:
```
user_id      1.408013e+08
recipe_id    1.600944e+05
rating       4.410802e+00
dtype: float64
```

1.5 Создайте серию из 10 случайных названий рецептов.

In [192…
```
recipes = np.random.choice(list(data_recipes["name"]), size = 10)
random_recipes = pd.Series(recipes)
random_recipes
```

Out[192]:
```
0    autumn sweet potato  or pumpkin  muffins
1         mustard grilled scandinavian salmon
2                             zucchini perini
3                   jan s curried lentil soup
4       blueberry cream cheese danish oatmeal
5         creamy chicken or turkey with pasta
6                     mustard glazed meatloaf
7            crab stuffed mushrooms en croute
8          spinach  bacon and mushroom quiche
9              gingered fish from the islands
dtype: object
```

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

In [195…
```
data_reviews.index = range(data_reviews.shape[0])

data_reviews.head()
```

Out[195]:

	user_id	recipe_id	date	rating	review
0	21752	57993	2003-05-01	5	Last week whole sides of frozen salmon fillet ...
1	431813	142201	2007-09-16	5	So simple and so tasty! I used a yellow capsi...
2	400708	252013	2008-01-10	4	Very nice breakfast HH, easy to make and yummy...
3	2001852463	404716	2017-12-11	5	These are a favorite for the holidays and so e...
4	95810	129396	2008-03-14	5	Excellent soup! The tomato flavor is just gre...

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

In [198…
```
data_recipes[(data_recipes['minutes'] <= 20) & (data_recipes['n_ingredients'] <= 5)]
```

Out[198]:

	name	id	minutes	contributor_id	submitted	n_steps	description	n_ingredients
28	quick biscuit bread	302399	20	213909	2008-05-06	11.0	this is a wonderful quick bread to make as an ...	5.0
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...	5.0
90	hawaiian sunrise mimosa	100837	5	58104	2004-09-29	4.0	pineapple mimosa was changed to hawaiian sunri...	3.0
91	tasty dish s banana pudding in 2 minutes	286484	2	47892	2008-02-13	NaN	"mmmm, i love bananas!" a --tasty dish-- origi...	4.0
94	1 minute meatballs	11361	13	4470	2001-09-03	NaN	this is a real short cut for cooks in a hurry....	2.0
...	...	...	...	...	...	...	...	...
29873	zip and steam red potatoes with butter and garlic	304922	13	724218	2008-05-27	9.0	i haven't tried this yet, but i am going to so...	5.0
29874	ziplock vanilla ice cream	74250	10	24386	2003-10-29	8.0	a fun thing for kids to do. may want to use mi...	3.0
29905	zucchini and corn with cheese	256177	15	305531	2007-09-29	4.0	from betty crocker fresh spring recipes. i lik...	5.0
29980	zucchini with jalapeno monterey jack	320622	10	305531	2008-08-20	3.0	simple and yummy!	3.0
29983	zucchini with serrano ham	162411	15	152500	2006-03-31	6.0	this dish is from tim malzer, a german chef wh...	5.0

2019 rows × 8 columns

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.

In [202…
```
data_recipes["submitted"] = pd.to_datetime(data_recipes["submitted"])

data_recipes.dtypes
```

Out[202]:
```
name                    object
id                       int64
minutes                  int64
contributor_id           int64
submitted       datetime64[ns]
n_steps                float64
description             object
n_ingredients          float64
dtype: object
```

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

In [214…
```
recipe_filtered = data_recipes[data_recipes["submitted"].dt.year <= 2010]
recipe_filtered.head()
```

Out[214]:

	name	id	minutes	contributor_id	submitted	n_steps	description	n_ingredients
0	george s at the cove black bean soup	44123	90	35193	2002-10-25	NaN	an original recipe created by chef scott meska...	18.0
1	healthy for them yogurt popsicles	67664	10	91970	2003-07-26	NaN	my children and their friends ask for my homem...	NaN
2	i can t believe it s spinach	38798	30	1533	2002-08-29	NaN	these were so go, it surprised even me.	8.0
3	italian gut busters	35173	45	22724	2002-07-27	NaN	my sister-in-law made these for us at a family...	NaN
4	love is in the air beef fondue sauces	84797	25	4470	2004-02-23	4.0	i think a fondue is a very romantic casual din...	NaN

Работа со строковыми данными в `pandas`

3.1 Добавьте в таблицу `recipes` столбец `description length`, в котором хранится длина описания рецепта из

столбца `description`.

In [243…
```
data_recipes["description_length"] = data_recipes["description"].fillna("").apply(len)
data_recipes[["description", "description_length"]].head()
```

Out[243]:

	description	description_length
0	an original recipe created by chef scott meska...	330
1	my children and their friends ask for my homem...	255
2	these were so go, it surprised even me.	39
3	my sister-in-law made these for us at a family...	154
4	i think a fondue is a very romantic casual din...	587

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

In [260…
```
data_recipes["name"] = data_recipes["name"].fillna("").apply(lambda x: x.title())
data_recipes.head()
```

Out[260]:

	name	id	minutes	contributor_id	submitted	n_steps	description	n_ingredients	description_length	name_word
0	George S At The Cove Black Bean Soup	44123	90	35193	2002-10-25	NaN	an original recipe created by chef scott meska...	18.0	330	
1	Healthy For Them Yogurt Popsicles	67664	10	91970	2003-07-26	NaN	my children and their friends ask for my homem...	NaN	255	
2	I Can T Believe It S Spinach	38798	30	1533	2002-08-29	NaN	these were so go, it surprised even me.	8.0	39	
3	Italian Gut Busters	35173	45	22724	2002-07-27	NaN	my sister-in-law made these for us at a family...	NaN	154	
4	Love Is In The Air Beef Fondue Sauces	84797	25	4470	2004-02-23	4.0	i think a fondue is a very romantic casual din...	NaN	587	

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.

In [258…
```
data_recipes["name_word_count"] = data_recipes["name"].fillna("").apply(lambda x: len(x.split()))

data_recipes[["name", "name_word_count"]].head()
```

Out[258]:

	name	name_word_count
0	george s at the cove black bean soup	8
1	healthy for them yogurt popsicles	5
2	i can t believe it s spinach	7
3	italian gut busters	3
4	love is in the air beef fondue sauces	8

Лабораторная работа №2.2

Группировки таблиц `pd.DataFrame`

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

In [278…
```
recipe_counts = data_recipes["contributor_id"].value_counts()
max_contributor = recipe_counts.idxmax()
```

```
max_recipes = recipe_counts.max()

print(f"Участник с ID {max_contributor} добавил максимальное количество рецептов: {max_recipes}")
```

```
Участник с ID 89831 добавил максимальное количество рецептов: 421
```

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.

In [541…
```
average_ratings = data_reviews.groupby('recipe_id')['rating'].mean()


average_ratings
```

Out[541]:
```
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
```

In [543…
```
sum((data_reviews[['recipe_id', 'rating']].groupby('recipe_id').mean() == 0)['rating'])
```

Out[543]:
```
660
```

4.3 Посчитайте количество рецептов с разбивкой по годам создания.

In [444…
```
data_recipes['year'] = data_recipes['submitted'].dt.year


recipes_by_year = data_recipes['year'].value_counts().sort_index()


pd.DataFrame(recipes_by_year)
```

Out[444]:

year	count
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

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`.

In [537…
```
merged_df = pd.merge(data_recipes, data_reviews, how='inner', left_on='id', right_on='recipe_id')

df = merged_df[['id', 'name', 'user_id', 'rating']]

# Поиск рецепта без отзывов
all_recipe_ids = set(data_recipes["id"])
reviewed_recipe_ids = set(df["id"])
unreviewed_recipe_ids = all_recipe_ids - reviewed_recipe_ids


unreviewed_recipe_id = next(iter(unreviewed_recipe_ids), None)


exists_in_filtered_df = unreviewed_recipe_id in df["id"].values

df.head()
```

Out[537]:

	id	name	user_id	rating
0	143615	Caramels	137302	5
1	143615	Caramels	267253	5
2	143615	Caramels	587361	5
3	143615	Caramels	17206	5
4	143615	Caramels	1488868	5

In [494…
```
unreviewed_recipe_id, exists_in_filtered_df
```

Out[494]:
```
(401411, False)
```

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`.

In [519…
```
review_counts = data_reviews["recipe_id"].value_counts().reset_index()
review_counts.columns = ["recipe_id", "review_count"]


recipes_review_counts = data_recipes[["id", "name"]].merge(review_counts, left_on="id", right_on="recipe_id", h


recipes_review_counts["review_count"] = recipes_review_counts["review_count"].fillna(0).astype(int)


recipe_without_reviews = recipes_review_counts[recipes_review_counts["review_count"] == 0]

recipes_review_counts.head()
```

Out[519]:

	id	name	recipe_id	review_count
0	143615	Caramels	143615.0	12
1	12957	Basbousa	12957.0	14
2	156521	Bushwhacker	156521.0	4
3	323195	Blackmoons	323195.0	1
4	119171	Mirepoix	119171.0	2

In [531…
```
recipe_without_reviews.sample(2)
```

Out[531]:

	id	name	recipe_id	review_count
17136	109432	Colorful Red Cabbage Salad	NaN	0
28582	341087	Judy S Hearty Split Pea And Ham Soup	NaN	0

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

In [517…
```
data_recipes["year"] = data_recipes["submitted"].dt.year


merged_df = data_reviews.merge(data_recipes, left_on="recipe_id", right_on="id", how="left")
```

```
avg_rating_by_year = merged_df.groupby("year")["rating"].mean()


lowest_rating_year = avg_rating_by_year.idxmin()
lowest_rating_value = avg_rating_by_year.min()

lowest_rating_year, lowest_rating_value
```

Out[517]: (2017, 2.75)

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.

In [504…
```
data_recipes = data_recipes.sort_values(by=["name_word_count"], ascending=True)
data_recipes.head()

data_recipes.to_csv("task_6_1.csv")
```

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.

In [533…
```
output_file = "task_6_2.xlsx"


with pd.ExcelWriter(output_file, engine="xlsxwriter") as writer:
    df.to_excel(writer, sheet_name="Рецепты с оценками", index=False)
    recipes_review_counts.to_excel(writer, sheet_name="Количество отзывов по рецептам", index=False)
```

[версия 2]

- Уточнены формулировки задач 1.1, 3.3, 4.2, 5.1, 5.2, 5.3