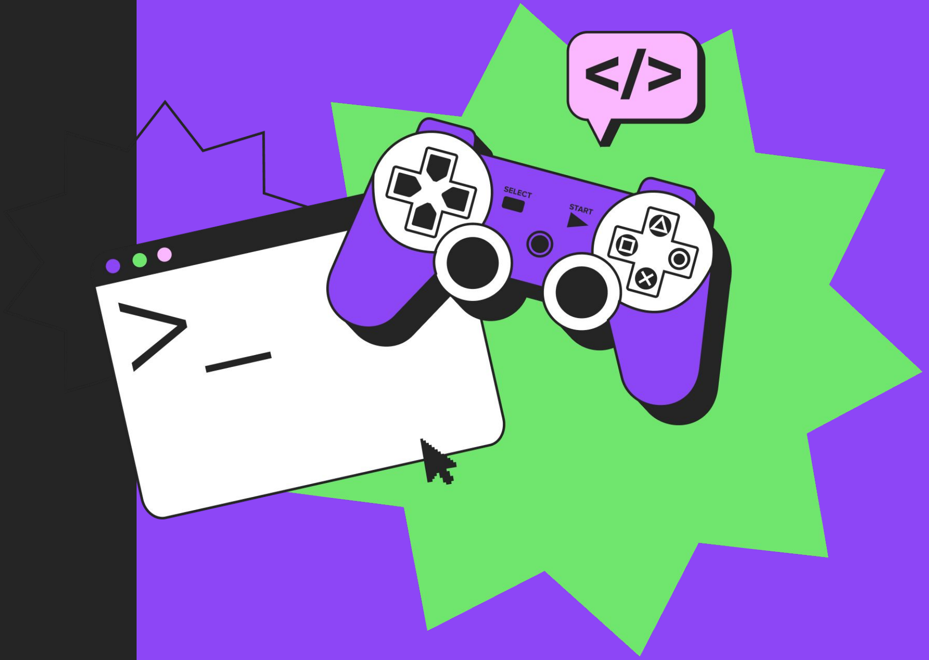
GeekBrains

Библиотеки NumPy и Pandas

Основы, которые нужно знать каждому Data Scientist’y

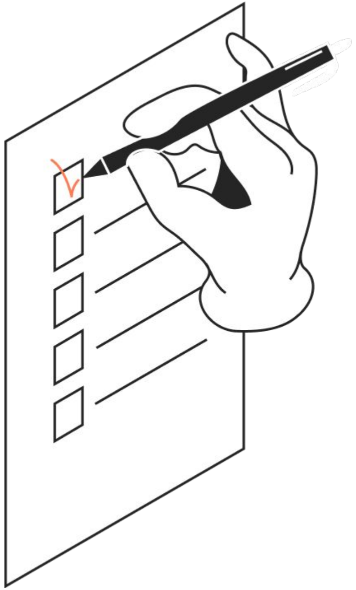


Разбор домашнего задания по теме «Знакомство с библиотеками»

→

Что будет на уроке сегодня

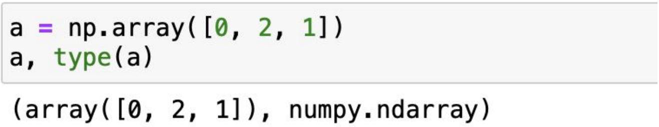
- 📌 Повторим NumPy и Pandas;
- 📌 Используем NumPy в бою;
- 📌 EDA с помощью Pandas.



NumPy. Невероятные возможности

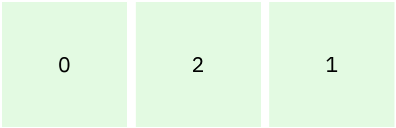
→

Одномерные массивы

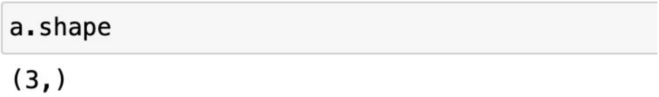


```
a = np.array([0, 2, 1])
a, type(a)
```

```
(array([0, 2, 1]), numpy.ndarray)
```

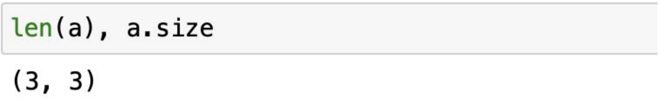


0 2 1



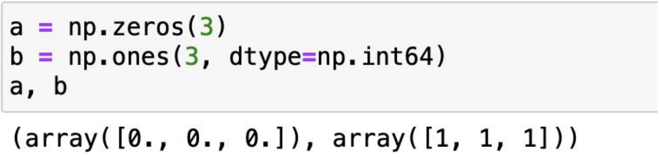
```
a.shape
```

```
(3,)
```



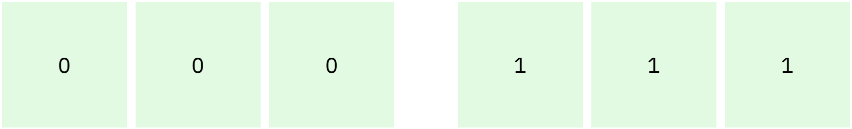
```
len(a), a.size
```

```
(3, 3)
```



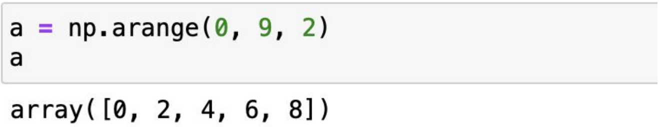
```
a = np.zeros(3)
b = np.ones(3, dtype=np.int64)
a, b
```

```
(array([0., 0., 0.]), array([1, 1, 1]))
```



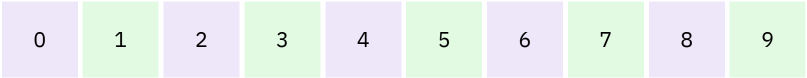
0 0 0

1 1 1



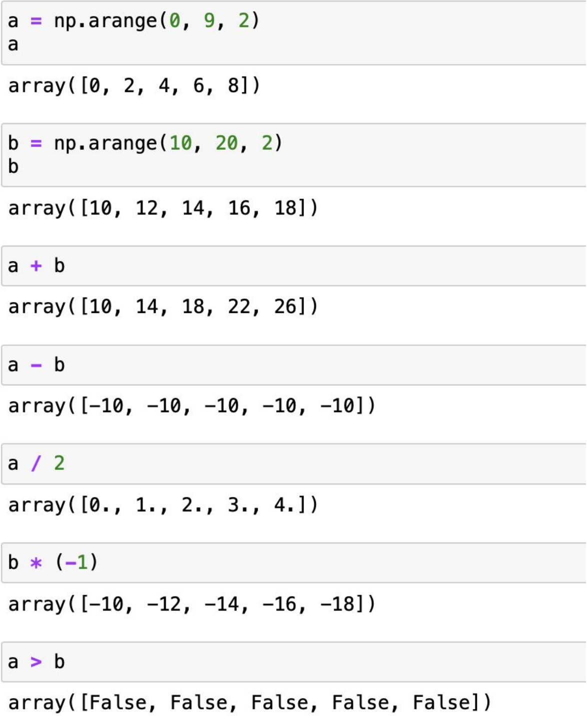
```
a = np.arange(0, 9, 2)
a
```

```
array([0, 2, 4, 6, 8])
```



0 1 2 3 4 5 6 7 8 9

Операции над одномерными массивами



```
a = np.arange(0, 9, 2)
a
```

```
array([0, 2, 4, 6, 8])
```

```
b = np.arange(10, 20, 2)
b
```

```
array([10, 12, 14, 16, 18])
```

```
a + b
```

```
array([10, 14, 18, 22, 26])
```

```
a - b
```

```
array([-10, -10, -10, -10, -10])
```

```
a / 2
```

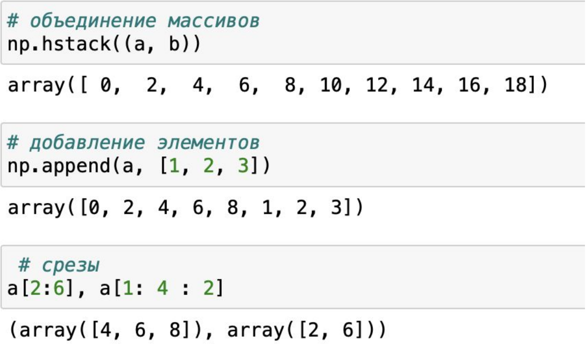
```
array([0., 1., 2., 3., 4.])
```

```
b * (-1)
```

```
array([-10, -12, -14, -16, -18])
```

```
a > b
```

```
array([False, False, False, False, False])
```



```
# объединение массивов
np.hstack((a, b))
```

```
array([ 0,  2,  4,  6,  8, 10, 12, 14, 16, 18])
```

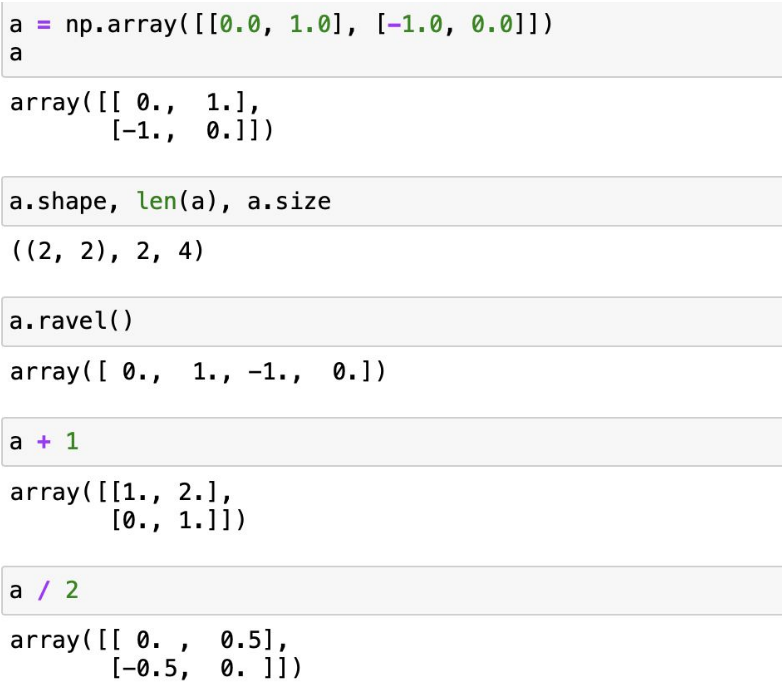
```
# добавление элементов
np.append(a, [1, 2, 3])
```

```
array([0, 2, 4, 6, 8, 1, 2, 3])
```

```
 # срезы
a[2:6], a[1: 4 : 2]
```

```
(array([4, 6, 8]), array([2, 6]))
```

Двумерные массивы



```
a = np.array([[0.0, 1.0], [-1.0, 0.0]])
a
```

```
array([[ 0.,  1.],
       [-1.,  0.]])
```

```
a.shape, len(a), a.size
```

```
((2, 2), 2, 4)
```

```
a.ravel()
```

```
array([ 0.,  1., -1.,  0.])
```

```
a + 1
```

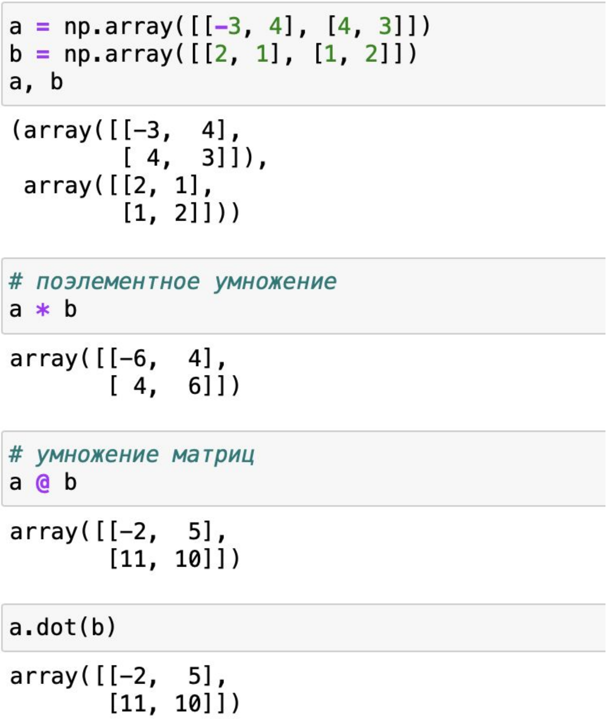
```
array([[1., 2.],
       [0., 1.]])
```

```
a / 2
```

```
array([[ 0. ,  0.5],
       [-0.5,  0. ]])
```

- Поддерживают все те же операции, что и одномерные (сложение с числом, деление на число и тд.);
- **.ravel()** — растягивание в одномерный массив.

Работа с матрицами



```
a = np.array([[-3, 4], [4, 3]])
b = np.array([[2, 1], [1, 2]])
a, b
```

```
(array([[-3,  4],
        [ 4,  3]]),
 array([[2, 1],
        [1, 2]]))
```

```
# поэлементное умножение
a * b
```

```
array([[-6,  4],
       [ 4,  6]])
```

```
# умножение матриц
a @ b
```

```
array([[-2,  5],
       [11, 10]])
```

```
a.dot(b)
```

```
array([[-2,  5],
       [11, 10]])
```

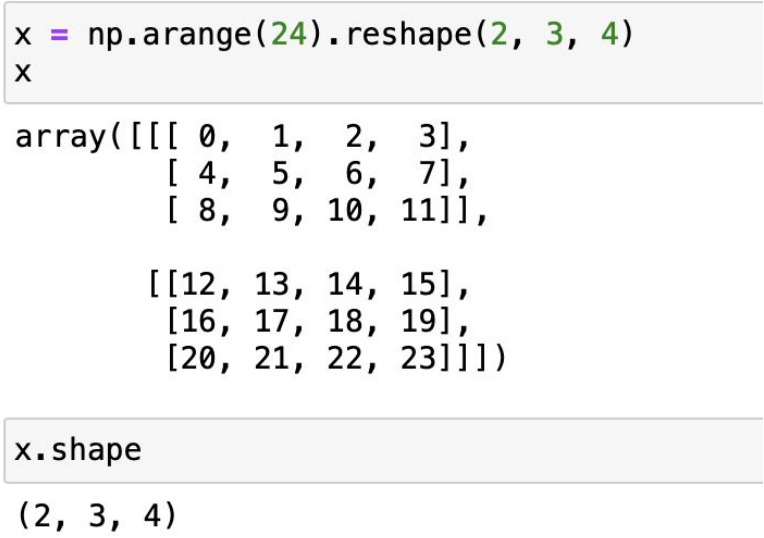
A

-3	4
4	3

B

2	1
1	2

Тензоры



```
x = np.arange(24).reshape(2, 3, 4)
x
```

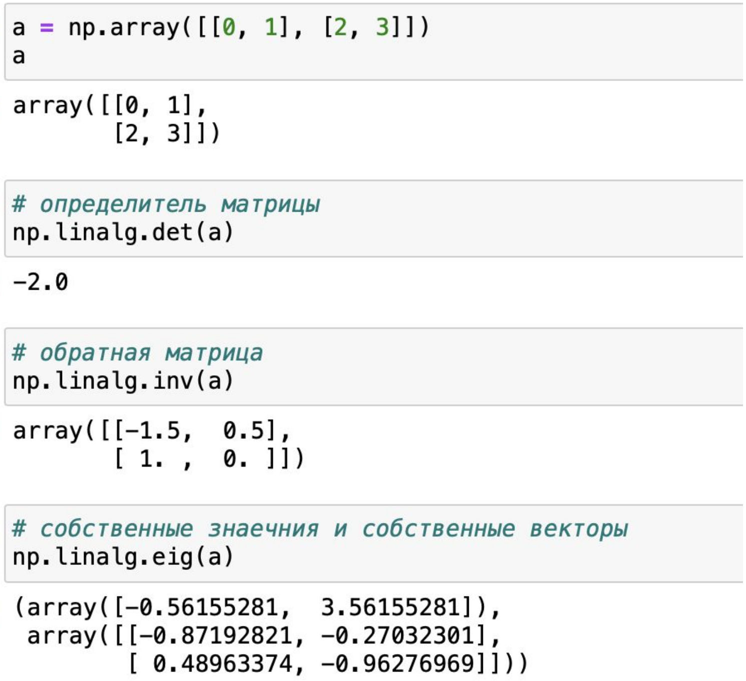
```
array([[[ 0,  1,  2,  3],
        [ 4,  5,  6,  7],
        [ 8,  9, 10, 11]],

       [[12, 13, 14, 15],
        [16, 17, 18, 19],
        [20, 21, 22, 23]]])
```

```
x.shape
```

```
(2, 3, 4)
```

- Тензор — многомерный массив;
- Поддерживает все те же операции, что и двумерный.

Линейная алгебра



```
a = np.array([[0, 1], [2, 3]])
a
```

```
array([[0, 1],
       [2, 3]])
```

```
# определитель матрицы
np.linalg.det(a)
```

```
-2.0
```

```
# обратная матрица
np.linalg.inv(a)
```

```
array([[-1.5,  0.5],
       [ 1. ,  0. ]])
```

```
# собственные знаечния и собственные векторы
np.linalg.eig(a)
```

```
(array([-0.56155281,  3.56155281]),
 array([[-0.87192821, -0.27032301],
        [ 0.48963374, -0.96276969]]))
```

- **.det** — определитель матрицы;
- **.inv** — обратная матрица;
- **.eig** — собственные векторы и собственные значения.

Pandas. EDA

Вопрос

Что такое EDA?

Ответ

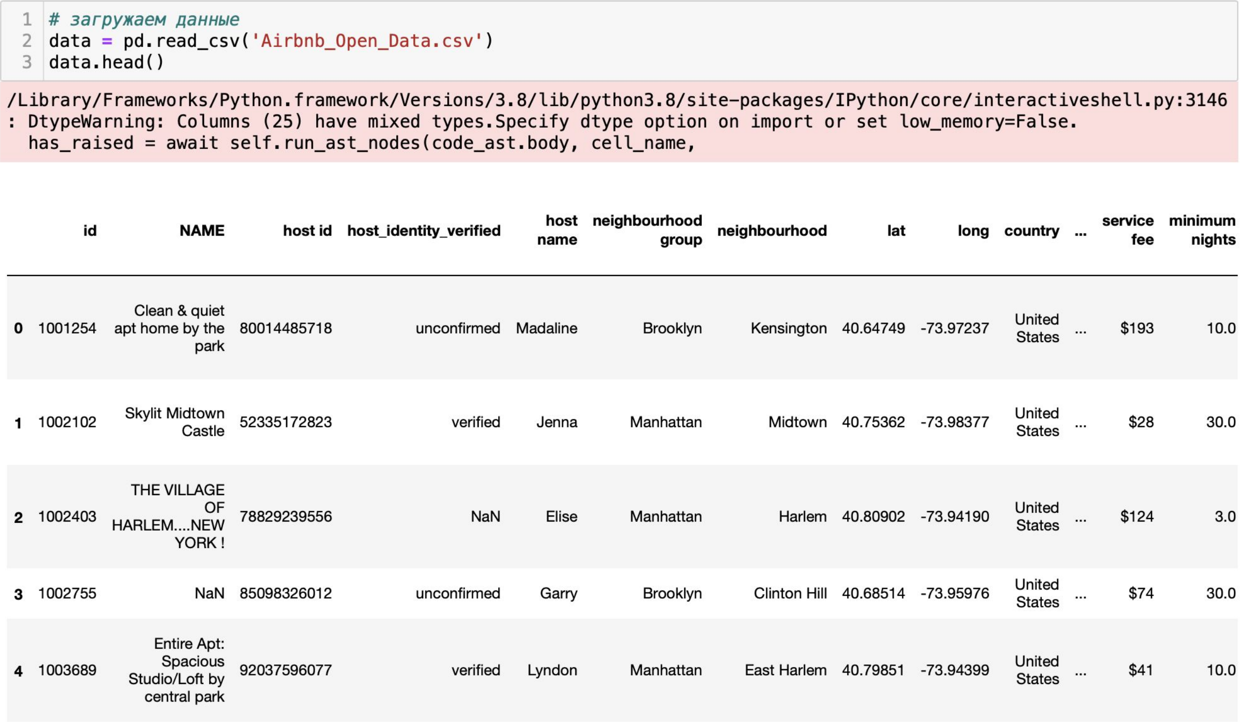
Exploratory Data Analysis

Exploratory Data Analysis

— разведочный анализ данных.

- «Погружение» в данные;
- Понимание структуры данных;
- Обнаружение аномалий и отклонений;
- Проверка основных гипотез и закономерностей.

Загрузка данных

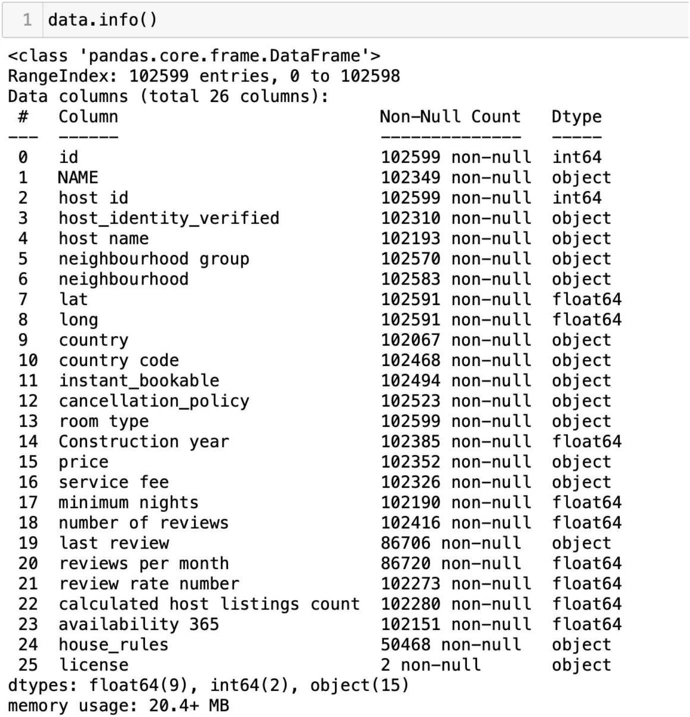


```
# загружаем данные
data = pd.read_csv('Airbnb_Open_Data.csv')
data.head()
```

```
/Library/Frameworks/Python.framework/Versions/3.8/lib/python3.8/site-packages/IPython/core/interactiveshell.py:3146
: DtypeWarning: Columns (25) have mixed types.Specify dtype option on import or set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
```

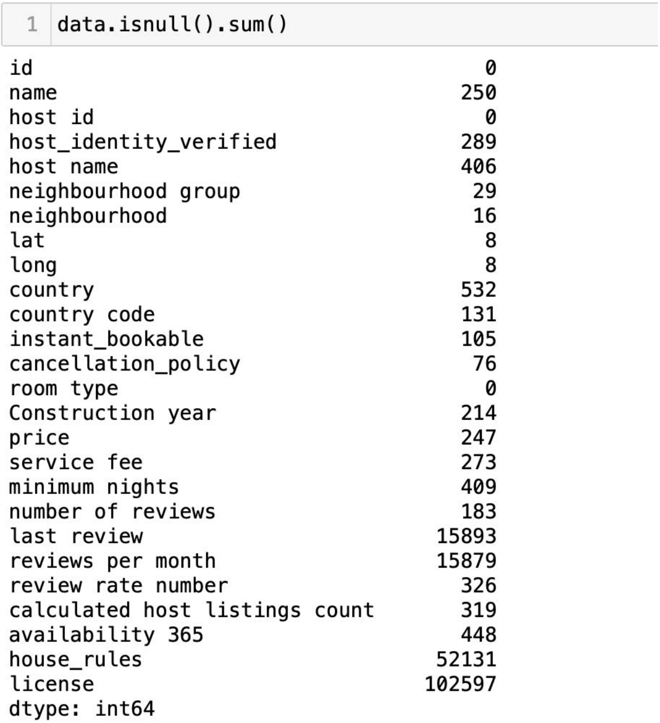
	id	NAME	host id	host_identity_verified	host name	neighbourhood group	neighbourhood	lat	long	country	...	service fee	minimum nights
0	1001254	Clean & quiet apt home by the park	80014485718	unconfirmed	Madaline	Brooklyn	Kensington	40.64749	-73.97237	United States	...	$193	10.0
1	1002102	Skylit Midtown Castle	52335172823	verified	Jenna	Manhattan	Midtown	40.75362	-73.98377	United States	...	$28	30.0
2	1002403	THE VILLAGE OF HARLEM....NEW YORK !	78829239556	NaN	Elise	Manhattan	Harlem	40.80902	-73.94190	United States	...	$124	3.0
3	1002755	NaN	85098326012	unconfirmed	Garry	Brooklyn	Clinton Hill	40.68514	-73.95976	United States	...	$74	30.0
4	1003689	Entire Apt: Spacious Studio/Loft by central park	92037596077	verified	Lyndon	Manhattan	East Harlem	40.79851	-73.94399	United States	...	$41	10.0

Описание датасета



```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 102599 entries, 0 to 102598
Data columns (total 26 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   id                              102599 non-null  int64
 1   NAME                            102349 non-null  object
 2   host id                         102599 non-null  int64
 3   host_identity_verified          102310 non-null  object
 4   host name                       102193 non-null  object
 5   neighbourhood group             102570 non-null  object
 6   neighbourhood                   102583 non-null  object
 7   lat                             102591 non-null  float64
 8   long                            102591 non-null  float64
 9   country                         102067 non-null  object
 10  country code                    102468 non-null  object
 11  instant_bookable                102494 non-null  object
 12  cancellation_policy             102523 non-null  object
 13  room type                       102599 non-null  object
 14  Construction year               102385 non-null  float64
 15  price                           102352 non-null  object
 16  service fee                     102326 non-null  object
 17  minimum nights                  102190 non-null  float64
 18  number of reviews               102416 non-null  float64
 19  last review                     86706 non-null   object
 20  reviews per month               86720 non-null   float64
 21  review rate number              102273 non-null  float64
 22  calculated host listings count  102280 non-null  float64
 23  availability 365                102151 non-null  float64
 24  house_rules                     50468 non-null   object
 25  license                         2 non-null       object
dtypes: float64(9), int64(2), object(15)
memory usage: 20.4+ MB
```



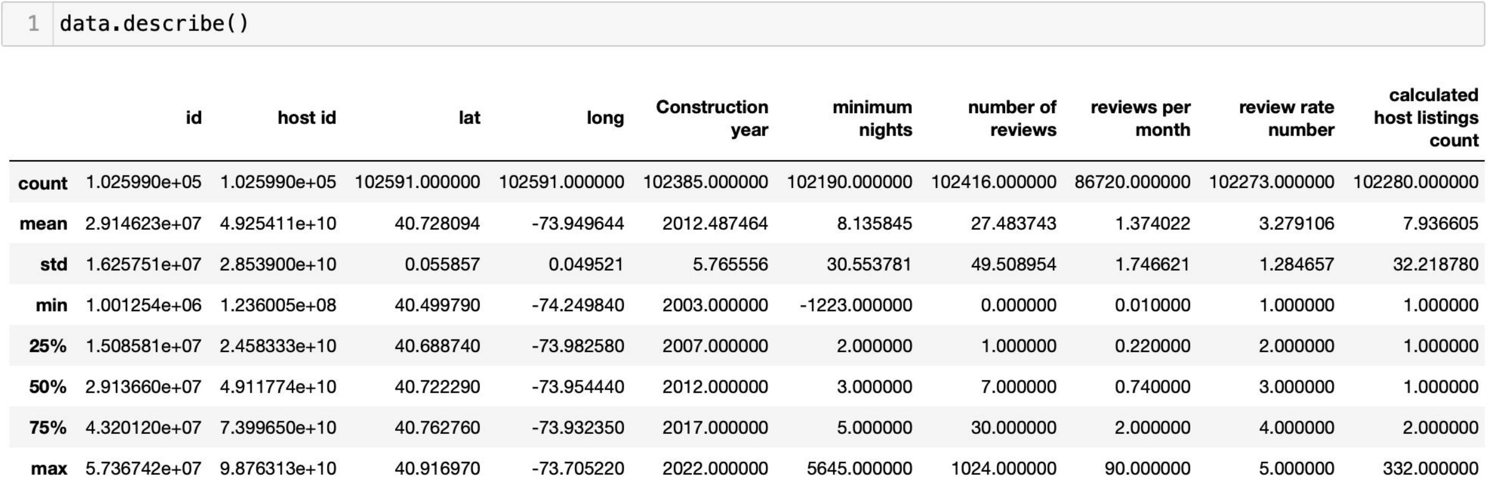
```
data.isnull().sum()
```

```
id                                     0
name                                 250
host id                                0
host_identity_verified               289
host name                            406
neighbourhood group                   29
neighbourhood                         16
lat                                    8
long                                   8
country                              532
country code                         131
instant_bookable                     105
cancellation_policy                   76
room type                              0
Construction year                    214
price                                247
service fee                          273
minimum nights                       409
number of reviews                    183
last review                        15893
reviews per month                  15879
review rate number                   326
calculated host listings count       319
availability 365                     448
house_rules                        52131
license                           102597
dtype: int64
```

Описание датасета

Основные статистики:

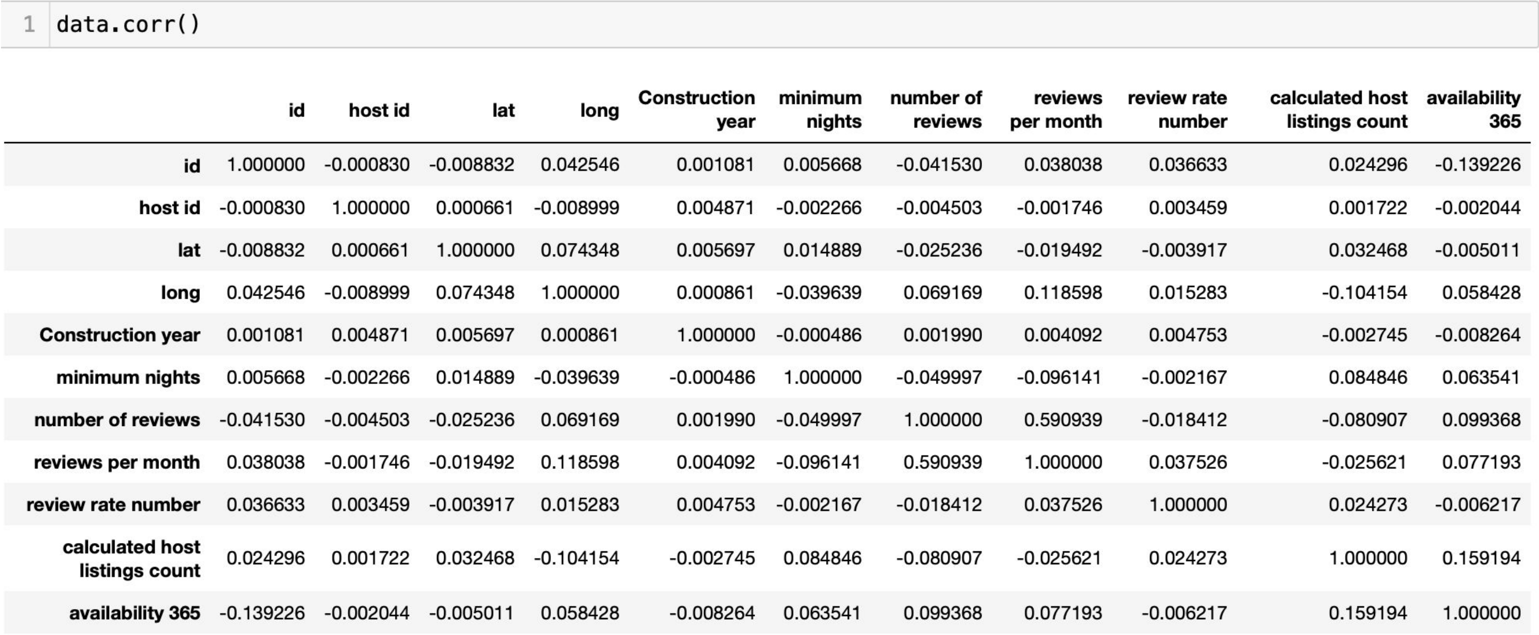
- Среднее (mean);
- Стандартное отклонение (std);
- Квартили (25/50/75);
- Минимум и максимум (min/max).



```
data.describe()
```

	id	host id	lat	long	Construction year	minimum nights	number of reviews	reviews per month	review rate number	calculated host listings count
count	1.025990e+05	1.025990e+05	102591.000000	102591.000000	102385.000000	102190.000000	102416.000000	86720.000000	102273.000000	102280.000000
mean	2.914623e+07	4.925411e+10	40.728094	-73.949644	2012.487464	8.135845	27.483743	1.374022	3.279106	7.936605
std	1.625751e+07	2.853900e+10	0.055857	0.049521	5.765556	30.553781	49.508954	1.746621	1.284657	32.218780
min	1.001254e+06	1.236005e+08	40.499790	-74.249840	2003.000000	-1223.000000	0.000000	0.010000	1.000000	1.000000
25%	1.508581e+07	2.458333e+10	40.688740	-73.982580	2007.000000	2.000000	1.000000	0.220000	2.000000	1.000000
50%	2.913660e+07	4.911774e+10	40.722290	-73.954440	2012.000000	3.000000	7.000000	0.740000	3.000000	1.000000
75%	4.320120e+07	7.399650e+10	40.762760	-73.932350	2017.000000	5.000000	30.000000	2.000000	4.000000	2.000000
max	5.736742e+07	9.876313e+10	40.916970	-73.705220	2022.000000	5645.000000	1024.000000	90.000000	5.000000	332.000000

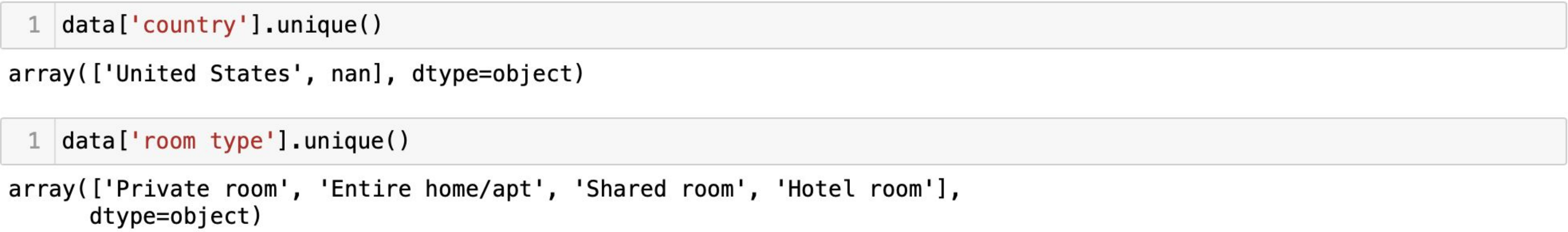
Корреляция признаков



```python
data.corr()
```

	id	host id	lat	long	Construction year	minimum nights	number of reviews	reviews per month	review rate number	calculated host listings count	availability 365
id	1.000000	-0.000830	-0.008832	0.042546	0.001081	0.005668	-0.041530	0.038038	0.036633	0.024296	-0.139226
host id	-0.000830	1.000000	0.000661	-0.008999	0.004871	-0.002266	-0.004503	-0.001746	0.003459	0.001722	-0.002044
lat	-0.008832	0.000661	1.000000	0.074348	0.005697	0.014889	-0.025236	-0.019492	-0.003917	0.032468	-0.005011
long	0.042546	-0.008999	0.074348	1.000000	0.000861	-0.039639	0.069169	0.118598	0.015283	-0.104154	0.058428
Construction year	0.001081	0.004871	0.005697	0.000861	1.000000	-0.000486	0.001990	0.004092	0.004753	-0.002745	-0.008264
minimum nights	0.005668	-0.002266	0.014889	-0.039639	-0.000486	1.000000	-0.049997	-0.096141	-0.002167	0.084846	0.063541
number of reviews	-0.041530	-0.004503	-0.025236	0.069169	0.001990	-0.049997	1.000000	0.590939	-0.018412	-0.080907	0.099368
reviews per month	0.038038	-0.001746	-0.019492	0.118598	0.004092	-0.096141	0.590939	1.000000	0.037526	-0.025621	0.077193
review rate number	0.036633	0.003459	-0.003917	0.015283	0.004753	-0.002167	-0.018412	0.037526	1.000000	0.024273	-0.006217
calculated host listings count	0.024296	0.001722	0.032468	-0.104154	-0.002745	0.084846	-0.080907	-0.025621	0.024273	1.000000	0.159194
availability 365	-0.139226	-0.002044	-0.005011	0.058428	-0.008264	0.063541	0.099368	0.077193	-0.006217	0.159194	1.000000

Уникальные значения в столбце



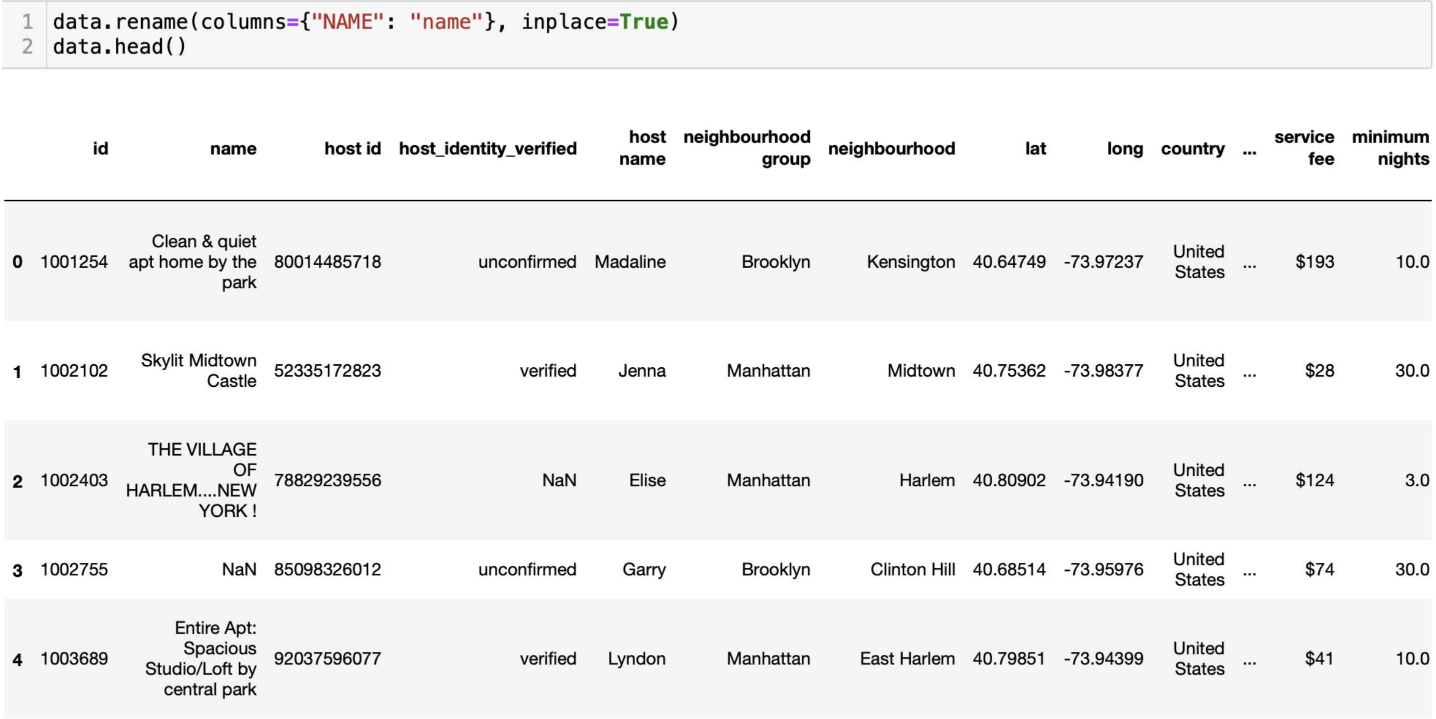
```
data['country'].unique()
```

```
array(['United States', nan], dtype=object)
```

```
data['room type'].unique()
```

```
array(['Private room', 'Entire home/apt', 'Shared room', 'Hotel room'],
      dtype=object)
```

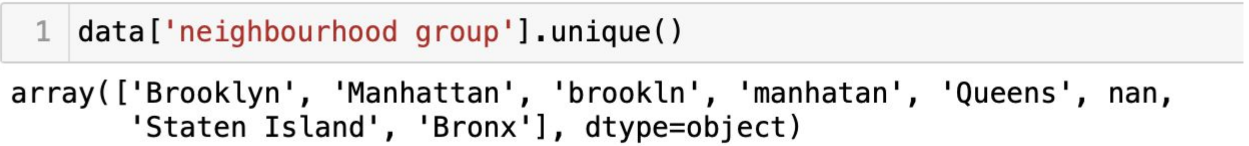
Переименование столбца



```python
data.rename(columns={"NAME": "name"}, inplace=True)
data.head()
```

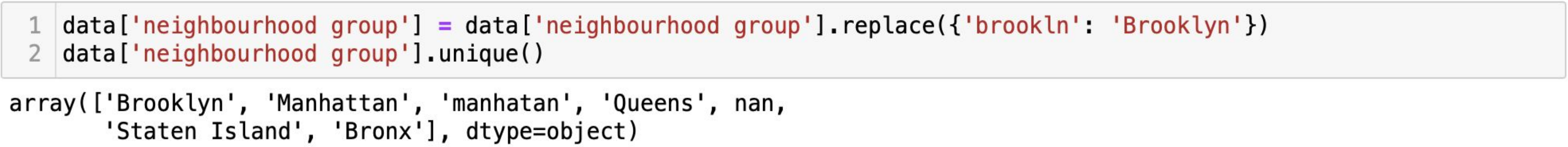
	id	name	host id	host_identity_verified	host name	neighbourhood group	neighbourhood	lat	long	country	...	service fee	minimum nights
0	1001254	Clean & quiet apt home by the park	80014485718	unconfirmed	Madaline	Brooklyn	Kensington	40.64749	-73.97237	United States	...	$193	10.0
1	1002102	Skylit Midtown Castle	52335172823	verified	Jenna	Manhattan	Midtown	40.75362	-73.98377	United States	...	$28	30.0
2	1002403	THE VILLAGE OF HARLEM....NEW YORK !	78829239556	NaN	Elise	Manhattan	Harlem	40.80902	-73.94190	United States	...	$124	3.0
3	1002755	NaN	85098326012	unconfirmed	Garry	Brooklyn	Clinton Hill	40.68514	-73.95976	United States	...	$74	30.0
4	1003689	Entire Apt: Spacious Studio/Loft by central park	92037596077	verified	Lyndon	Manhattan	East Harlem	40.79851	-73.94399	United States	...	$41	10.0

Переименование элементов в столбце



```python
data['neighbourhood group'].unique()
```

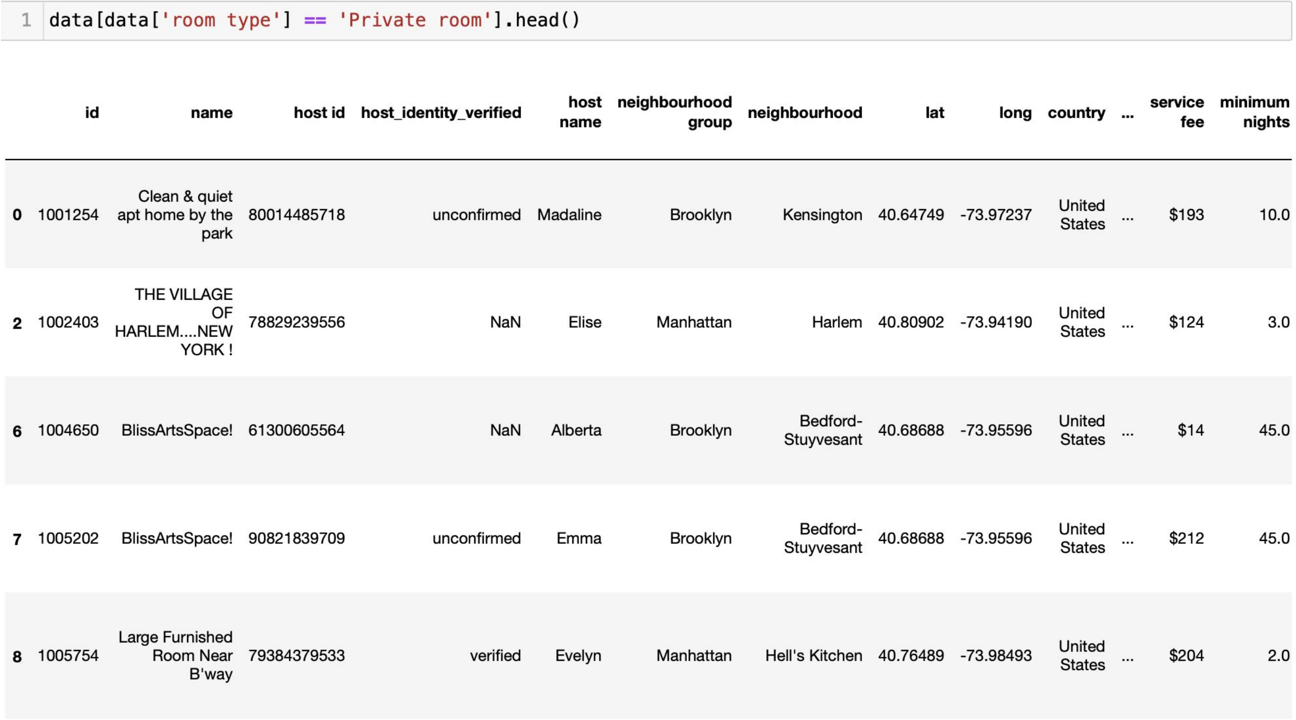
```
array(['Brooklyn', 'Manhattan', 'brookln', 'manhatan', 'Queens', nan,
       'Staten Island', 'Bronx'], dtype=object)
```



```python
data['neighbourhood group'] = data['neighbourhood group'].replace({'brookln': 'Brooklyn'})
data['neighbourhood group'].unique()
```

```
array(['Brooklyn', 'Manhattan', 'manhatan', 'Queens', nan,
       'Staten Island', 'Bronx'], dtype=object)
```

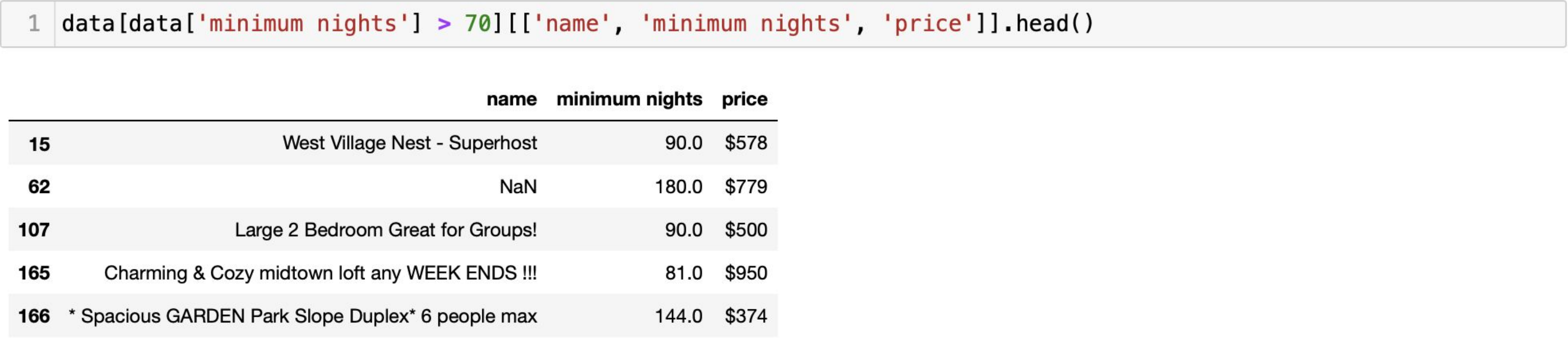
Фильтрация данных



```python
data[data['room type'] == 'Private room'].head()
```

	id	name	host id	host_identity_verified	host name	neighbourhood group	neighbourhood	lat	long	country	...	service fee	minimum nights
0	1001254	Clean & quiet apt home by the park	80014485718	unconfirmed	Madaline	Brooklyn	Kensington	40.64749	-73.97237	United States	...	$193	10.0
2	1002403	THE VILLAGE OF HARLEM....NEW YORK !	78829239556	NaN	Elise	Manhattan	Harlem	40.80902	-73.94190	United States	...	$124	3.0
6	1004650	BlissArtsSpace!	61300605564	NaN	Alberta	Brooklyn	Bedford-Stuyvesant	40.68688	-73.95596	United States	...	$14	45.0
7	1005202	BlissArtsSpace!	90821839709	unconfirmed	Emma	Brooklyn	Bedford-Stuyvesant	40.68688	-73.95596	United States	...	$212	45.0
8	1005754	Large Furnished Room Near B'way	79384379533	verified	Evelyn	Manhattan	Hell's Kitchen	40.76489	-73.98493	United States	...	$204	2.0

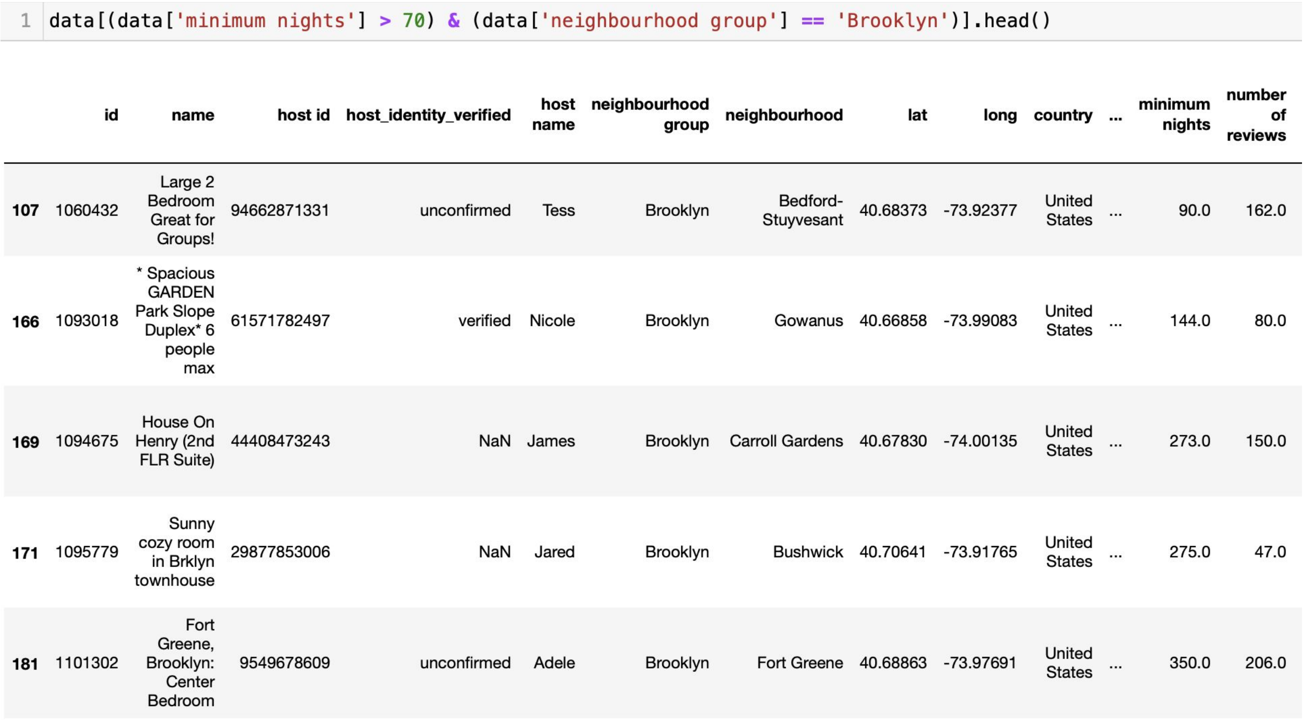
Фильтрация данных + отбор столбцов



```python
data[data['minimum nights'] > 70][['name', 'minimum nights', 'price']].head()
```

	name	minimum nights	price
15	West Village Nest - Superhost	90.0	$578
62	NaN	180.0	$779
107	Large 2 Bedroom Great for Groups!	90.0	$500
165	Charming & Cozy midtown loft any WEEK ENDS !!!	81.0	$950
166	* Spacious GARDEN Park Slope Duplex* 6 people max	144.0	$374

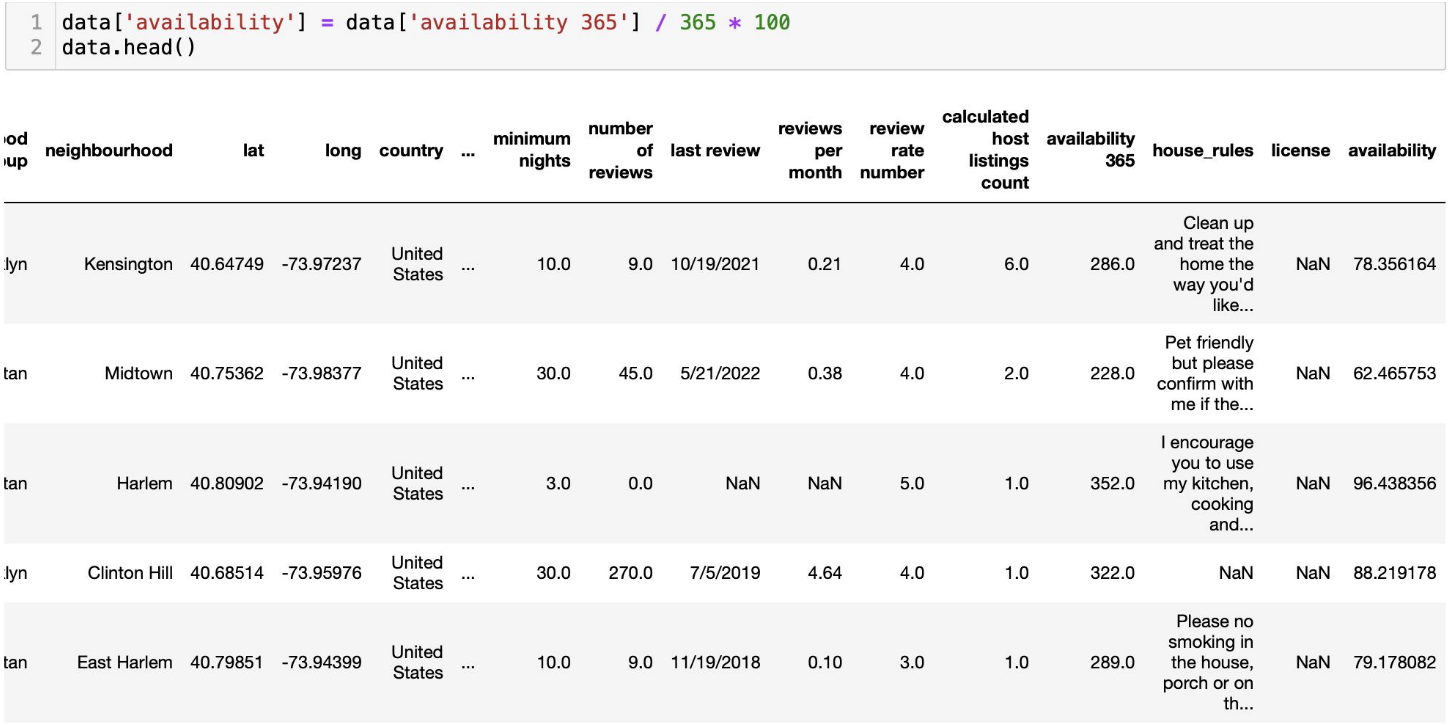
Фильтрация данных по нескольким признакам



```python
data[(data['minimum nights'] > 70) & (data['neighbourhood group'] == 'Brooklyn')].head()
```

	id	name	host id	host_identity_verified	host name	neighbourhood group	neighbourhood	lat	long	country	...	minimum nights	number of reviews
107	1060432	Large 2 Bedroom Great for Groups!	94662871331	unconfirmed	Tess	Brooklyn	Bedford-Stuyvesant	40.68373	-73.92377	United States	...	90.0	162.0
166	1093018	* Spacious GARDEN Park Slope Duplex* 6 people max	61571782497	verified	Nicole	Brooklyn	Gowanus	40.66858	-73.99083	United States	...	144.0	80.0
169	1094675	House On Henry (2nd FLR Suite)	44408473243	NaN	James	Brooklyn	Carroll Gardens	40.67830	-74.00135	United States	...	273.0	150.0
171	1095779	Sunny cozy room in Brklyn townhouse	29877853006	NaN	Jared	Brooklyn	Bushwick	40.70641	-73.91765	United States	...	275.0	47.0
181	1101302	Fort Greene, Brooklyn: Center Bedroom	9549678609	unconfirmed	Adele	Brooklyn	Fort Greene	40.68863	-73.97691	United States	...	350.0	206.0

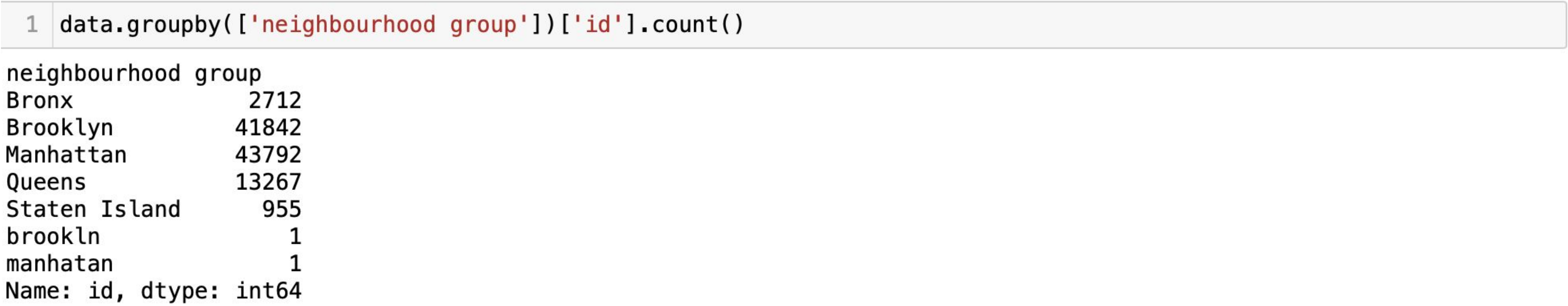
Добавление нового столбца



```python
data['availability'] = data['availability 365'] / 365 * 100
data.head()
```

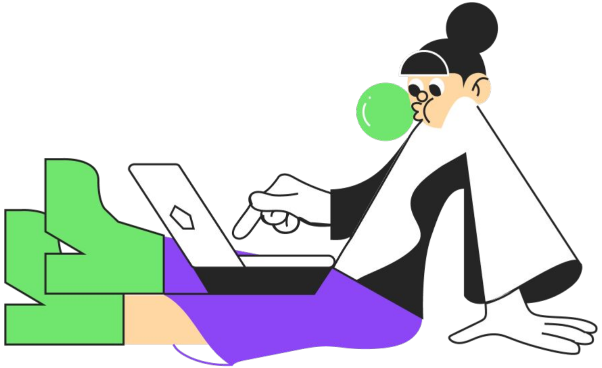
od up	neighbourhood	lat	long	country	...	minimum nights	number of reviews	last review	reviews per month	review rate number	calculated host listings count	availability 365	house_rules	license	availability
lyn	Kensington	40.64749	-73.97237	United States	...	10.0	9.0	10/19/2021	0.21	4.0	6.0	286.0	Clean up and treat the home the way you'd like...	NaN	78.356164
tan	Midtown	40.75362	-73.98377	United States	...	30.0	45.0	5/21/2022	0.38	4.0	2.0	228.0	Pet friendly but please confirm with me if the...	NaN	62.465753
tan	Harlem	40.80902	-73.94190	United States	...	3.0	0.0	NaN	NaN	5.0	1.0	352.0	I encourage you to use my kitchen, cooking and...	NaN	96.438356
lyn	Clinton Hill	40.68514	-73.95976	United States	...	30.0	270.0	7/5/2019	4.64	4.0	1.0	322.0	NaN	NaN	88.219178
tan	East Harlem	40.79851	-73.94399	United States	...	10.0	9.0	11/19/2018	0.10	3.0	1.0	289.0	Please no smoking in the house, porch or on th...	NaN	79.178082

Группировка данных по признакам



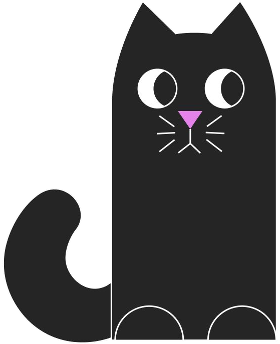
```python
data.groupby(['neighbourhood group'])['id'].count()
```

```
neighbourhood group
Bronx               2712
Brooklyn           41842
Manhattan          43792
Queens             13267
Staten Island        955
brookln                1
manhatan               1
Name: id, dtype: int64
```



Практическое задание

Ищите практическое задание в notebook с уроком.



Что мы узнали сегодня

- 📌 Повторили NumPy и Pandas;
- 📌 Использовали NumPy в бою;
- 📌 EDA с помощью Pandas.

Вопросы?

Вопросы?

Вопросы?