Работа с пропусками

Операции со столбцами датафрейма.

Уникальные значения столбца и их количество

```
data['device_type'].value_counts()
```

Арифметические операции со столбцами

```
# столбцы должны иметь числовой тип
purchases['total'] = purchases['first'] + purchases['repeated']
purchases['difference'] = purchases['first'] - purchases['repeated']
data['total_spent'] = data['clicks'] * data['click_price']
data['conversion'] = data['orders'] / data['visits']
```

Различные агрегации для столбца

```
# применяется только для числовых столбцов
data['clicks_per_day'].sum() # сумма кликов за все дни
data['visits_per_day'].min() # минимальное количество посещений за день
data['daily_revenue'].max() # максимальная прибыль за день
data['salary'].mean() # средняя зарплата
data['salary'].median() # медианная зарплата

# также можно подсчитать количество значений в столбце без учёта пропусков
data['clicks_per_day'].count()
```

Применение разных агрегаций к разным столбцам группировки

```
# предположим, в таблицу с результатами факультетов Хогвартса добавили возраст ученико
hogwarts_points.groupby('faculty_name').agg({'points': ['sum'], 'age': ['min', 'max']}
```

В результате для каждого факультета получим не только сумму баллов, но также минимальный и максимальный возраст студентов.

Работа с пропусками

Поиск пропусков

```
# подсчитает количество пропусков в столбце 'email' датафрейма logs
logs['email'].isna().sum()

# вернёт все строки датафрейма logs с пропусками в столбце 'email'
logs[logs['email'].isna()]
```

Заполнение пропусков

```
# замена пропусков в столбце 'email' датафрейма logs на пустую строку
logs['email'] = logs['email'].fillna('')

# замена пропусков в столбце 'age' датафрейма metrica на среднее значение в этом столбце
metrica['age'] = metrica['age'].fillna(value=metrica['age'].mean())

# замена пропусков во всех столбцах датафрейма
logs = logs.fillna('')
```

Заполнение пропусков в количественных переменных по группам

```
# заменяем все пропуски в столбце 'time' на среднее значение
# среди такого же типа устройств
for d in metrica['device_type'].unique():
    metrica.loc[(metrica['device_type'] == d) & (metrica['time'].isna()), 'time'] = \
                    metrica.loc[(metrica['device_type'] == d), 'time'].mean()
```

Глоссарий

User ID — уникальный номер, который присваивается посетителю сайта, чтобы отличать его от остальных и запоминать его поведение.

NaN (англ. not a number) — специальное значение типа float, которое используется, если результат вычисления не может или не должен быть представлен как конкретное число, или попросту неизвестен.

Работа с пропусками

Категориальная переменная — переменная, которая принимает одно значение из ограниченного набора.

Количественная переменная — переменная, которая принимает любое числовое значение в диапазоне.

Логическая (булева) переменная — переменная, принимающая значение True **(истина) или False **(ложь).

Словарь — структура данных, которая хранит набор элементов в виде пары «ключ — значение».