PYTHON PANDAS YORDAMIDA MA’LUMOTLARNI TOZALASH VA OLDINDAN QAYTA ISHLASH

Saidabonu Shokirova
Chirchiq Davlat Pedagogika Universiteti

Annotatsiya: *Ushbu maqola Pythonning Pandas kutubxonasi yordamida ma'lumotlarni tozalash va almashtirishning amaliy usullarini koʻrib chiqadi. Ob-havo ma'lumotlari va talabalar baholari misollarida yetishmayotgan qiymatlarni boshqarish, notoʻgʻri qiymatlarni almashtirish, matnlarni qayta ishlash va kategoriyali ma'lumotlarni raqamli shaklga oʻtkazish kabi jarayonlar tushuntiriladi. Maqola Pandasning replace, regex va boshqa funksiyalaridan foydalanishni amaliy misollar bilan ochib beradi.*

Kalit soʻzlar: *Pandas, Python, Ma'lumotlarni tozalash, Data preprocessing, NaN qiymatlari, Replace funksiyasi, Kategoriyali ma'lumotlar, Matnli ma'lumotlarni raqamlashtirish*

KIRISH

Ma'lumotlarni tozalash va qayta ishlash ma'lumotlar tahlilining muhim bosqichidir. Pythonning Pandas kutubxonasi bu jarayonni samarali va osonlashtirilgan tarzda amalga oshirish uchun keng qamrovli vositalarni taqdim etadi[1]. Ushbu maqolada ob-havo ma'lumotlari va talabalar baholari misollarida Pandas yordamida ma'lumotlarni tozalash, notoʻgʻri qiymatlarni almashtirish va matnlarni qayta ishlash usullari amaliy misollar bilan koʻrib chiqiladi.

Ma'lumotlarni qayta ishlashni boshlashdan oldin kerakli kutubxonalarni import qilamiz. NumPy (raqamli hisoblash uchun) va Pandas (ma'lumotlar bilan ishlash uchun). np va pd qisqartmalari kodni ixchamlashtirish uchun ishlatiladi.

```
import numpy as np
import pandas as pd
```

Ob-havo ma'lumotlari weather_data.csv faylidan pd.read_csv() funksiyasi yordamida oʻqiladi va DataFrame shaklida saqlanadi. data oʻzgaruvchisi ma'lumotlarni konsolda koʻrsatadi[2].

```
data = pd.read_csv('weather_data.csv')
data
```

Natija:

	DAY	TEMPERATURE	WINDSPEED	EVENT
	1/1/2017	32	6	Rain
	1/2/2017	-99999	7	Sunny
	1/3/2017	28	-99999	Snow

	DAY	TEMPERATURE	WINDSPEED	EVENT
	1/4/2017	-99999	7	0
	1/5/2017	32	-99999	Rain
	1/6/2017	31	2	Sunny
	1/6/2017	34	5	0

Shape atributi DataFrame’ning qator va ustunlar sonini qaytaradi, bu ma'lumotlar to‘plamining umumiy tuzilishini tushunishga yordam beradi.

Data.shape()

Natija (7,4)

info() metodi har bir ustunning ma'lumot turini (masalan, int64, object) va yetishmayotgan qiymatlar sonini ko‘rsatadi. Bu ma'lumotlarni tozalashdan oldin tuzilishni tushunish uchun muhim.

```
data.info()
Natija
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7 entries, 0 to 6
Data columns (total 4 columns):
day          7 non-null object
temperature  7 non-null int64
windspeed    7 non-null int64
event        7 non-null object
dtypes: int64(2), object(2)
memory usage: 352.0+ bytes
```

day va event ustunlari matnli (object), temperature va windspeed ustunlari esa butun son (int64) turiga ega. Hozircha yetishmayotgan qiymatlar yo‘q.

-99999 qiymati noto‘g‘ri deb hisoblanadi va uni np.NaN bilan almashtirish uchun replace() funksiyasi ishlatiladi[3]. Bu yetishmayotgan ma'lumotlarni aniqlashni osonlashtiradi.

data = data.replace(-99999,value=np.NaN)

data

Natija:

	DAY	TEMPERATURE	WINDSPEED	EVENT
0	1/1/2017	32.0	6.0	Rain
1	1/2/2017	NaN	7.0	Sunny
2	1/3/2017	28.0	NaN	Snow
3	1/4/2017	NaN	7.0	0
4	1/5/2017	32.0	NaN	Rain
5	1/6/2017	31.0	2.0	Sunny
6	1/6/2017	34.0	5.0	0

NaN qiymatlari qo‘shilgandan so‘ng, temperature va windspeed ustunlarining ma'lumot turlari int64 dan float64 ga o‘zgardi, chunki NaN suzuvchi nuqtali sonlar bilan ishlaydi.

```
data.info()
Natija:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7 entries, 0 to 6
Data columns (total 4 columns):
day          7 non-null object
temperature    5 non-null float64
windspeed      5 non-null float64
event          7 non-null object
dtypes: float64(2), object(2)
memory usage: 352.0+ bytes
```

Qiymatlar ro‘yxatini yagona qiymat bilan almashtirish

Izoh: 32.0 va 7.0 qiymatlari 99 ga almashtiriladi. replace() funksiyasi ro‘yxatdagi bir nechta qiymatni bir vaqtning o‘zida o‘zgartirish imkonini beradi.

```
data = data.replace([32.0,7.0],value=99)
data
```

Natija:

	DAY	TEMPERATURE	WINDSPEED	EVENT
	1/1/2017	99.0	6.0	Rain
	1/2/2017	NaN	99.0	Sunny
	1/3/2017	28.0	NaN	Snow
	1/4/2017	NaN	99.0	0
	1/5/2017	99.0	NaN	Rain
	1/6/2017	31.0	2.0	Sunny
	1/6/2017	34.0	5.0	0

Ustun bo‘yicha almashtirish

Izoh: Har bir ustundagi ma'lum qiymatlarni alohida almashtirish uchun lug‘at yordamida replace() funksiyasi ishlatiladi. temperature ustunidagi 99.0 ni 100 ga, windspeed ustunidagi NaN ni 100 ga, event ustunidagi 0 ni 100 ga almashtiramiz.

```
data.replace({'temperature':99.0,'windspeed':np.nan,'event':'0'},100)
```

Natija:

day temperature windspeed event

```
0  01/1/2017      100.0      6.0   Rain
1  11/2/2017       NaN      99.0  Sunny
2  21/3/2017       28.0     100.0   Snow
3  31/4/2017       NaN      99.0    100
4  41/5/2017      100.0     100.0   Rain
5  51/6/2017       31.0       2.0  Sunny
6  61/6/2017       34.0       5.0    100
```

NaN va 0 qiymatlari lug‘at yordamida bir vaqtning o‘zida almashtiriladi. NaN 69 ga, 0 esa Sunny ga o‘zgartiriladi.

```
data = data.replace({np.nan:69,'0':'Sunny'})
data
```

weather2.csv faylida temperature va windspeed ustunlarida F va mph kabi matnli birliklar mavjud. replace() funksiyasi va regex=True opsiyasi yordamida bu birliklar olib tashlanadi.

```
data = pd.read_csv('weather2.csv')
data
```

Talabalar va ularning baholari haqidagi ma'lumotlar lug‘at sifatida yaratiladi va pd.DataFrame() yordamida DataFrame’ga aylantiriladi.

```
d = {'score':['exceptional','average','good','poor','average','exceptional'],
'student':['Karan','Arpit','Varun','Robin','Akshay','Ankush']}
data = pd.DataFrame(d)
data
```

Natija:

```
score  student
0  exceptional    Karan
1      average    Arpit
2         good    Varun
3         poor    Robin
4      average   Akshay
5  exceptional   Ankush
```

Matnli baholar (poor, average, good, exceptional) raqamli shaklga (1, 2, 3, 4) aylantiriladi. Bu kategoriyali ma'lumotlarni tahlil uchun moslashtiradi.

```
data = data.replace(['poor','average','good','exceptional'],[1,2,3,4])
data
```

Natija:

```
score  student
0      4    Karan
1      2    Arpit
2      3    Varun
3      1    Robin
4      2   Akshay
```

5 4 Ankush

Xulosa. Pandas kutubxonasi ma'lumotlarni tozalash va oldindan qayta ishlash jarayonini sezilarli darajada soddalashtiradi. Ushbu maqolada ob-havo va talabalar ma'lumotlari misollarida replace(), regex va boshqa Pandas funksiyalarining qo‘llanilishi ko‘rsatildi. Bu usullar noto‘g‘ri qiymatlarni boshqarish, matnli ma'lumotlarni tozalash va kategoriyali ma'lumotlarni raqamli shaklga o‘tkazishda muhim ahamiyatga ega. Pandas yordamida ma'lumotlarni samarali tozalash va transformatsiya qilish real dunyo ma'lumotlari bilan ishlashda aniq va ishonchli tahlil natijalarini ta'minlaydi.

FOYDALANILGAN ADABIYOTLAR:

1. Pandas Documentation – https://pandas.pydata.org/docs/
2. Tojiddinov, A., Gulsumoy, N., Muntazam, H., & Tojimamatov, I. (2023). BIG DATA. Journal of Integrated Education and Research, 2(3), 35-42.
3. VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media.
4. Kazil, J., & Jarmul, K. (2016). Data Wrangling with Python: Tips and Tools to Make Your Life Easier. O'Reilly Media.
5. https://realpython.com
6. Esanovna D. B. Modern Teaching Aids and Technical Equipment in Modern Educational Institutions //International Journal of Innovative Analyses and Emerging Technology. – Т. 2. – №. 6.