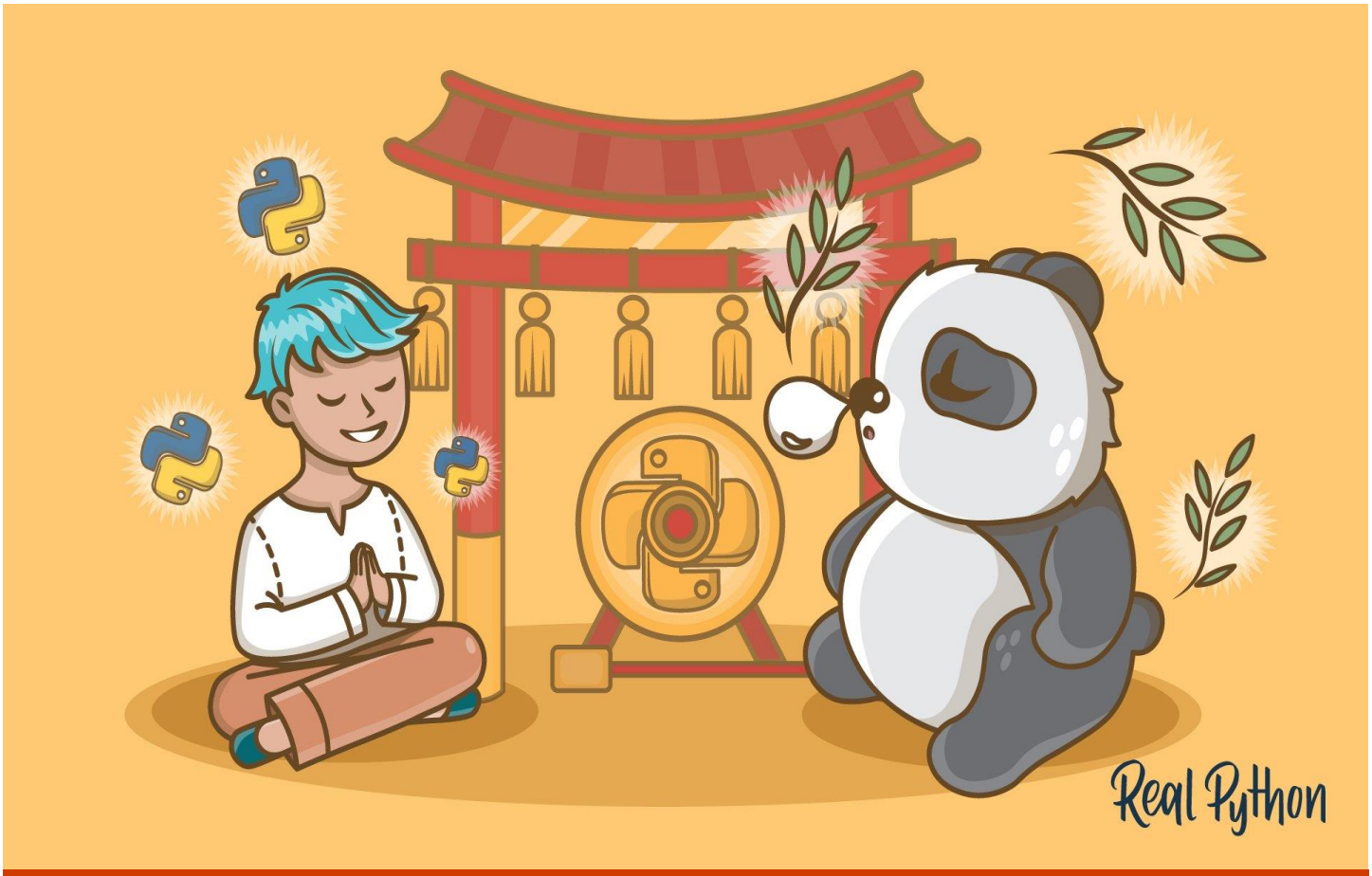

دوره جامع پایتون در یادگیری ماشین

قسمت ششم، دیتا های خالی یا گم شده

Pandas[3]

Datalk.ir
Created by : Ali Arabshahi
Contact us : Linkedin.com/in/mrAliArabshahi

دیتاهای خالی یا گم شده (Missing Values)

در این جلسه به این می پردازیم که چطوری میشه با دیتا هایی که در جداول مون درج نشدن سر و کله بزنیم 🤦

طبق معمول، کتابخونه های نامپای و پانداس رو فراخونی می کنیم:

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

حالا به دیتا فریم می سازیم که بعضی از مقادیرش **خالی** یا **نال** باشن.

```
df = pd.DataFrame({'A':[1,2,np.nan],
                   'B':[5,np.nan,np.nan],
                   'C':[1,2,3]})
```



```
df
```

	A	B	C
0	1.0	5.0	1
1	2.0	NaN	2
2	NaN	NaN	3

توسط دستور زیر میتونیم سطر هایی که دارای مقادر گم شده هستند، مشاهده کنیم که البته مشاهده کردن معمولا خیلی هم به کارمون نمی یاد! باید بدونیم باهاشون چی کار کنیم. دستور زیر رو من خودم نمی دونستم. رفتم و در گوگل جمله زیر رو سرچ کردم 🤠

```
how to find null rows pandas!
```

و کلی سایت که از راهنمایی یکیشون استفاده کردم؛ باور کنین به عنوان یه دیتا ساینتیست خیلی خیلی **حرفه ای**! اصلا لازم نیست همه چیز رو بلد باشین؛ مگر این که یکم آماتور باشین! 🤦

```
df[df.isnull().any(axis=1)]
```

	A	B	C
1	2.0	NaN	2
2	NaN	NaN	3

I. حذف داده های گم شده

یکی از رویکرد هایی که می تونیم نسبت به این نوع از دیتا ها داشته باشیم اینه که کلا حذفشون کنیم، یا به قول معروف صورت مسئله رو پاک کنیم! 😎 لزوما بهترین راه نیست اما قطعا **ساده ترین** راهه:

```
df.dropna()
```

	A	B	C
0	1.0	5.0	1

در حالت پیشفرض اکسیس (axis) روی صفر قرار داره و این یعنی هر سطری که حتی یک مقدار نال داشت، سر به نیست کن. 🤐 در مثال پایین اکسیس رو برابر یک میزاریم و بنابرین این سانسور روی هر ستونی اتفاق می افته که یک مقدار گم شده داره.

```
df.dropna(axis=1)
```

	C
0	1
1	2
2	3

خب شاید رویکرد بالا یکم منطقی نباشه! فکر کنین ما صد تا ستون داریم و فقط یکی از مقادیر اون نال هست! روش بالا می یاد و کل ردیف رو حذف می کنه! برای رو به رو شدن با این چالش، پانداس راه حل زیر رو پیشنهاد می کنه که می آیم و یک حد یا آستانه ای برای اون تعریف می کنیم. پانداس جان! ❤️‍🩹 اگر در ردیفی بیشتر از دو مقدارِ گم شده قرار داشت، اون رو حذف کن، اما اگر یکی بود بی خیال شو!

```
df.dropna(thresh=2)
```

	A	B	C
0	1.0	5.0	1
1	2.0	NaN	2

II. جایگزینی مقادیر گم شده

رویکرد دوم، که بیشتر هم مرسوم هست اینه که داده گم شده مون رو با یک مقدار دیگه جایگزین کنیم. این مقدار می تونه یک متن باشه، مشابه زیر:


```
df.fillna(value='FILL VALUE')
```

	A	B	C
0	1	5	1
1	2	FILL VALUE	2
2	FILL VALUE	FILL VALUE	3

و یا می تونه یک عدد باشه، فرض کنین ستون اِی! درآمد افراد رو نشون می ده و ما به پانداس می گیم اگر کسی فراموش کرد درآمدش رو وارد کنه، بیا و میانگین کل درآمد ها رو محاسبه کن و سپس با اون مقدار فراموش شده جایگزین کن:

```
df['A'].fillna(value=df['A'].mean())
```

```
0    1.0
1    2.0
2    1.5
Name: A, dtype: float64
```

پس در این جلسه یاد گرفتیم چطوری میشه با مقادیری که در دیتا بیس مون برابر **هیچی** هستن، مواجه بشیم و باهاشون کنار بیایم. تا جلسه بعد

خیلی مراقب خودتون باشین

///😊\\\