Knowledge Discovery and Data Mining

Lab 4 Data Cleaning II Dates, Encoding Types and Remove Duplications

Xuan Song
Songx@sustech.edu.cn

Topics

1. Play with Datetime type in pandas Dataframe
2. Understand different kinds of character encodings
3. Remove duplicate records

Dates

Let’s start by printing out the date column, shall we?

```
0    01/02/1965
1    01/04/1965
2    01/05/1965
3    01/08/1965
4    01/09/1965
Name: Date, dtype: object
```

We can clearly see that a string like “01/02/1965” to be a date. In python, this is called a “datetime” type. However, when we read the csv file, this structure is not automatically maintained, and instead, we just get the default “object” type.

```
dtype('O')
```

Date

We will use the **pandas.to_datetime()** function to convert the object type column into datetime type column.

```
0   1965-01-02 00:00:00+00:00
1   1965-01-04 00:00:00+00:00
2   1965-01-05 00:00:00+00:00
3   1965-01-08 00:00:00+00:00
4   1965-01-09 00:00:00+00:00
Name: Date_parsed, dtype: datetime64[ns, UTC]
```

If you encounter problems when converting datetime, refer to these 2 following links:

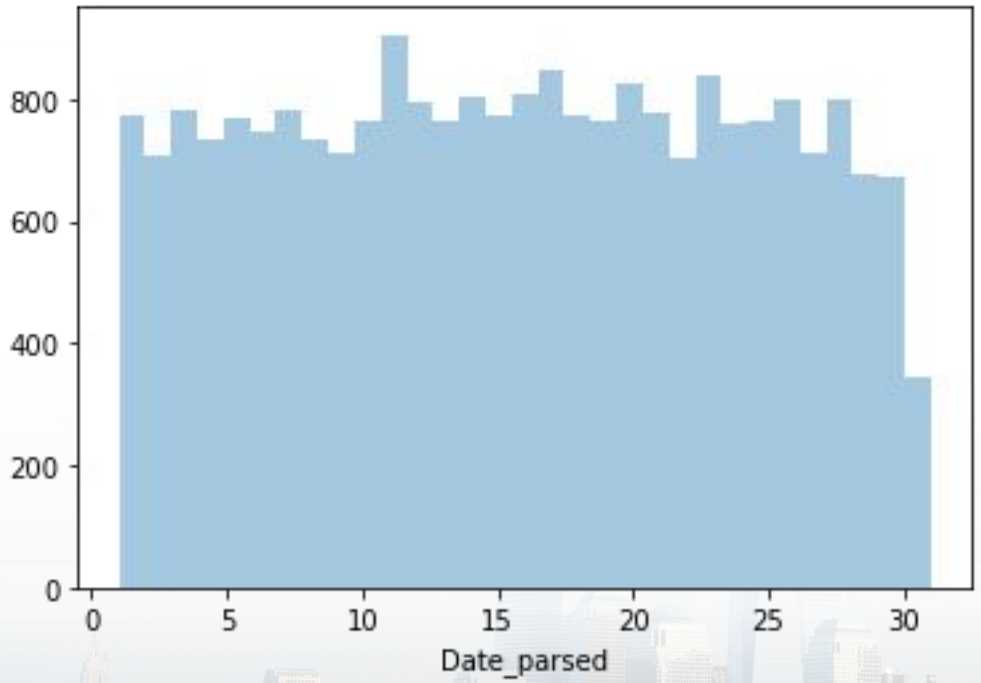
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.to_datetime.html

https://docs.python.org/zh-cn/3/library/datetime.html#strftime-and-strptime-format-codes

Date

Cool, we get the date column in format “datetime”, now what?

We can start extracting the day information from the column and plot out the day distribution.



Hint:
Use **pandas.Series.dt.day()** to extract each datetime from the column.
Use **seaborn.distplot()** to make the plot.

```
day_of_month_earthquakes = earthquakes['Date_parsed'].dt.day
```

```
sns.distplot(day_of_month_earthquakes, kde=False, bins=31)
```

Date – Lab Exercise

Make a day plot AND a week-of-day plot of both data:

Data 1: landslide_catalog.csv

Data 2: volcano_database.csv

Character Encoding

Sometimes, the file you try to read in might not be the convenient encoding type (the default standard encoding is type ‘utf-8’).

```
UnicodeDecodeError: 'utf-8' codec can't decode byte 0x99 in position 11: invalid start byte
```

But let’s first play with the character codings first: Try encoding and decoding different symbols to ASCII and see what happens. I'd recommend $, #, 你好 and न म स॒ ते but feel free to try other characters as well.

Character Encoding

```python
# start with a string
before = "This is the euro symbol: €"

# encode it to a different encoding, replacing characters that raise errors
after = before.encode("ascii", errors = "replace")

# convert it back to utf-8
print(after.decode("ascii"))

# We've lost the original underlying byte string! It's been
# replaced with the underlying byte string for the unknown character :(
```

```
This is the euro symbol: ?
```

https://docs.python.org/zh-cn/3/library/codecs.html#standard-encodings

Character Encoding

One way to find out which character encoding your file contains is by utilizing the python chardet function.

```
# look at the first ten thousand bytes to guess the character encoding
with open("ks-projects-201612.csv", 'rb') as rawdata:
    result = chardet.detect(rawdata.read(10000))

# check what the character encoding might be
print(result)
```

```
{'encoding': 'Windows-1252', 'confidence': 0.73, 'language': ''}
```

Now we have our initial guess to how to correctly decode the file!

Character Encoding – Lab Exercise

Successfully read in these two data:

Data 1: ks-projects-201801.csv

Data 2: PoliceKillingsUS.csv

Duplication – Lab Exercise

This one is relatively easy, just use the pandas default **drop_duplicates()** function.

Now, calculate the percentage of data retained after deduplication:

Data to use: Reviews.csv 

Hint: use len(your_dataframe) to get the length.

Class Work

As explained above in the 3 sessions.

No extra challenge this week, but you are more than welcome to play around with the given datasets.

Starting next week we will begin model training ☺

Homework 1

Homework 1 is also up!

Make sure you check out
Blackboard and start working on it!

End of Lab 4