datacleaning_NYU

December 7, 2022

1 Data Cleaning for NYU

[ ]:
```
import pandas as pd
import numpy as np
from datetime import datetime
date_format = "%Y-%m-%d"
```

[ ]:
```
raw_data = pd.read_csv('Datasets/2000_01_01_2022_11_01_NYU.csv')
```

[ ]:
```
raw_data.head()
```

[ ]:
```
     name    datetime  tempmax  tempmin  temp  feelslikemax  feelslikemin  \
0  10012  2000-01-01      9.5      1.3   4.6           7.0          -1.9
1  10012  2000-01-02     14.1      4.3   9.6          14.1           4.0
2  10012  2000-01-03     16.5      9.5  13.3          16.5           8.2
3  10012  2000-01-04     18.9      9.1  12.4          18.9           7.5
4  10012  2000-01-05      8.5     -1.0   3.2           5.4          -5.8

   feelslike   dew  humidity  …  solarenergy  uvindex  severerisk  \
0        2.0   3.7      94.1  …          NaN      NaN         NaN
1        9.3   7.9      89.2  …          NaN      NaN         NaN
2       13.2  11.0      86.7  …          NaN      NaN         NaN
3       12.1  11.1      91.7  …          NaN      NaN         NaN
4       -2.0  -5.1      55.5  …          NaN      NaN         NaN

                sunrise               sunset  moonphase  \
0  2000-01-01T07:20:08  2000-01-01T16:38:45       0.89
1  2000-01-02T07:20:15  2000-01-02T16:39:35       0.93
2  2000-01-03T07:20:20  2000-01-03T16:40:27       0.96
3  2000-01-04T07:20:22  2000-01-04T16:41:21       0.98
4  2000-01-05T07:20:22  2000-01-05T16:42:16       1.00

              conditions                                        description  \
0       Partially cloudy                Partly cloudy throughout the day.
1               Overcast                 Cloudy skies throughout the day.
2               Overcast                 Cloudy skies throughout the day.
3         Rain, Overcast        Cloudy skies throughout the day with rain.
```

```
4  Rain, Partially cloudy  Partly cloudy throughout the day with early mo…

                icon                                          stations
0  partly-cloudy-day  72503794745,72502014734,74486094789,72503014732
1             cloudy  72503794745,72502014734,74486094789,72503014732
2             cloudy  72503794745,72502014734,74486094789,72503014732
3               rain  72503794745,72502014734,74486094789,72503014732
4               rain  72503794745,72502014734,74486094789,72503014732

[5 rows x 33 columns]
```

[ ]:
```
print(raw_data.conditions.unique())
```

```
['Partially cloudy' 'Overcast' 'Rain, Overcast' 'Rain, Partially cloudy'
 'Snow, Rain, Partially cloudy' 'Clear' 'Snow, Rain, Overcast'
 'Snow, Partially cloudy' 'Snow, Overcast' 'Rain' 'Snow' 'Snow, Rain'
 'Snow, Rain, Freezing Drizzle/Freezing Rain, Overcast'
 'Rain, Freezing Drizzle/Freezing Rain, Ice, Partially cloudy'
 'Snow, Rain, Ice, Overcast']
```

[ ]:
```
print(raw_data.icon.unique())
```

```
['partly-cloudy-day' 'cloudy' 'rain' 'wind' 'snow' 'clear-day']
```

[ ]:
```
all_seasons = raw_data[['datetime', 'conditions']]
```

[ ]:
```
all_seasons.head()
```

[ ]:
```
     datetime              conditions
0  2000-01-01        Partially cloudy
1  2000-01-02                Overcast
2  2000-01-03                Overcast
3  2000-01-04          Rain, Overcast
4  2000-01-05  Rain, Partially cloudy
```

[ ]:
```
# all_seasons['datetime'] = [datetime.strptime(dt, date_format) for dt in␣
↪all_seasons['datetime']]
# all_seasons['quarter'] = [dt.quarter for dt in all_seasons['datetime']]
```

[ ]:
```
# all_seasons.quarter.unique()
```

[ ]:
```
# winter = all_seasons[all_seasons.quarter == 1]
# spring = all_seasons[all_seasons.quarter == 2]
# summer = all_seasons[all_seasons.quarter == 3]
# fall = all_seasons[all_seasons.quarter == 4]
```

[ ]:
```
all_seasons.to_csv('Datasets/all_seasons_NYU.csv', encoding='utf-8',␣
↪index=False)
# winter.to_csv('winter.csv', encoding='utf-8', index=False)
# spring.to_csv('spring.csv', encoding='utf-8', index=False)
# summer.to_csv('summer.csv', encoding='utf-8', index=False)
# fall.to_csv('fall.csv', encoding='utf-8', index=False)
```