datacleaning_UCLA

December 7, 2022

1 Data cleaning for UCLA

[ ]:
```
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_UCLA.csv')
```

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

[ ]:
```
    name    datetime  tempmax  tempmin  temp  feelslikemax  feelslikemin  \
0  90024  2000-01-01     13.9      8.1  10.7          13.9           7.7
1  90024  2000-01-02     15.7      8.1  12.4          15.7           6.6
2  90024  2000-01-03     18.6      6.3  11.9          18.6           5.5
3  90024  2000-01-04     19.6      8.2  13.3          19.6           8.2
4  90024  2000-01-05     21.2      7.3  13.9          21.2           7.3

   feelslike  dew  humidity  …  solarenergy  uvindex  severerisk  \
0       10.6  5.5      70.8  …          NaN      NaN         NaN
1       12.2 -0.4      44.4  …          NaN      NaN         NaN
2       11.8 -0.2      46.1  …          NaN      NaN         NaN
3       13.2  1.7      48.3  …          NaN      NaN         NaN
4       13.8  2.2      47.9  …          NaN      NaN         NaN

                sunrise               sunset  moonphase        conditions  \
0  2000-01-01T06:59:26  2000-01-01T16:55:04       0.89  Partially cloudy
1  2000-01-02T06:59:37  2000-01-02T16:55:49       0.93             Clear
2  2000-01-03T06:59:47  2000-01-03T16:56:36       0.96             Clear
3  2000-01-04T06:59:55  2000-01-04T16:57:23       0.98             Clear
4  2000-01-05T07:00:01  2000-01-05T16:58:12       1.00             Clear

                           description               icon  \
0    Partly cloudy throughout the day.  partly-cloudy-day
1  Clear conditions throughout the day.          clear-day
2  Clear conditions throughout the day.          clear-day
3  Clear conditions throughout the day.          clear-day
```

```
4  Clear conditions throughout the day.          clear-day

                              stations
0  72295023174,72287493134,72297023129
1  72295023174,72287493134,72297023129
2  72295023174,72287493134,72297023129
3  72295023174,72287493134,72297023129
4  72295023174,72287493134,72297023129

[5 rows x 33 columns]
```

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

```
['Partially cloudy' 'Clear' 'Rain, Partially cloudy' 'Rain'
 'Rain, Overcast' 'Overcast']
```

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

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

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

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

[ ]:

	datetime	conditions
0	2000-01-01	Partially cloudy
1	2000-01-02	Clear
2	2000-01-03	Clear
3	2000-01-04	Clear
4	2000-01-05	Clear
5	2000-01-06	Clear
6	2000-01-07	Partially cloudy
7	2000-01-08	Partially cloudy
8	2000-01-09	Partially cloudy
9	2000-01-10	Partially cloudy
10	2000-01-11	Partially cloudy
11	2000-01-12	Partially cloudy
12	2000-01-13	Rain, Partially cloudy
13	2000-01-14	Partially cloudy
14	2000-01-15	Partially cloudy
15	2000-01-16	Rain, Partially cloudy
16	2000-01-17	Rain, Partially cloudy
17	2000-01-18	Partially cloudy
18	2000-01-19	Partially cloudy
19	2000-01-20	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']]
```

```
/var/folders/6m/88dwrhnx7m3cybxwl1p0rtq40000gn/T/ipykernel_80665/3833035098.py:1
: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  all_seasons['datetime'] = [datetime.strptime(dt, date_format) for dt in
all_seasons['datetime']]
/var/folders/6m/88dwrhnx7m3cybxwl1p0rtq40000gn/T/ipykernel_80665/3833035098.py:2
: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  all_seasons['quarter'] = [dt.quarter for dt in all_seasons['datetime']]
```

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

[ ]:
```
array([1, 2, 3, 4])
```

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