datacleaning_UCSD

December 7, 2022

1 Data Cleaning for UCSD

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

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

```
[ ]:    name    datetime  tempmax  tempmin  temp  feelslikemax  feelslikemin  \
     0  92093  2000-01-01     14.7     10.1  12.4          14.7          10.1
     1  92093  2000-01-02     15.1      9.6  12.9          15.1           9.6
     2  92093  2000-01-03     18.9      6.6  12.1          18.9           6.6
     3  92093  2000-01-04     18.5      7.8  13.1          18.5           6.8
     4  92093  2000-01-05     17.1      7.3  11.7          17.1           7.3

        feelslike  dew  humidity  …  solarenergy  uvindex  severerisk  \
     0       12.4  8.1      75.6  …          NaN      NaN         NaN
     1       12.9  5.8      63.1  …          NaN      NaN         NaN
     2       12.1  2.5      54.9  …          NaN      NaN         NaN
     3       12.9  3.6      54.8  …          NaN      NaN         NaN
     4       11.7  6.0      70.0  …          NaN      NaN         NaN

                    sunrise               sunset  moonphase  \
     0  2000-01-01T06:50:28  2000-01-01T16:51:18       0.89
     1  2000-01-02T06:50:40  2000-01-02T16:52:02       0.93
     2  2000-01-03T06:50:51  2000-01-03T16:52:48       0.96
     3  2000-01-04T06:51:00  2000-01-04T16:53:35       0.98
     4  2000-01-05T06:51:07  2000-01-05T16:54:22       1.00

                    conditions                                      description  \
     0  Rain, Partially cloudy  Partly cloudy throughout the day with late aft…
     1        Partially cloudy                Partly cloudy throughout the day.
     2                   Clear             Clear conditions throughout the day.
     3                   Clear             Clear conditions throughout the day.
```

```
4                   Clear              Clear conditions throughout the day.

                icon                  stations
0               rain  72290693112,72290023188
1  partly-cloudy-day  72290693112,72290023188
2          clear-day  72290693112,72290023188
3          clear-day  72290693112,72290023188
4          clear-day  72290693112,72290023188

[5 rows x 33 columns]
```

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

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

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

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

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

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

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

```
[ ]: # 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_69670/3833035098.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_69670/3833035098.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('Datasets/all_seasons_UCSD.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)
```