

```
In [97]: import pandas as pd
import datetime

use_average = True
use_predict = True

output_name = 'average.csv' if use_average else 'normal.csv'

output_name2020 = '2020average.csv' if use_average else '2020normal.csv'
```

```
In [98]: def read_csv(file):
df = pd.read_csv(file)
del df['Station Name']
del df['Climate ID']
del df['Longitude (x)']
del df['Latitude (y)']
del df['Year']
del df['Month']
del df['Day']
del df['Time']

df.columns = ['date', 'temp', 'dew', 'rel', 'wdir', 'wspeed', 'stn']
df['date'] = pd.to_datetime(df['date'])
return df
```

```
In [99]: if use_predict:
df1 = read_csv('HALIFAX DOCKYARD.csv')
df2 = read_csv('HALIFAX KOOTENAY.csv')
else:
df1 = read_csv('2020-HALIFAX DOCKYARD.csv')
df2 = read_csv('2020-HALIFAX KOOTENAY.csv')

if use_average:
data = {'date': df1['date'],
'temp': (df1['temp'] + df2['temp'])/2.0,
'dew': (df1['dew'] + df2['dew'])/2.0,
'rel': (df1['rel'] + df2['rel'])/2.0,
'wdir': (df1['wdir'] + df2['wdir'])/2.0,
'wspeed': (df1['wspeed'] + df2['wspeed'])/2.0,
'stn': (df1['stn'] + df2['stn'])/2.0,
}
else:
data = {'date': df1['date'],
's1_temp': df1['temp'],
's2_temp': df2['temp'],
's1_dew': df1['dew'],
's2_dew': df2['dew'],
's1_rel': df1['rel'],
's2_rel': df2['rel'],
's1_wdir': df1['wdir'],
's2_wdir': df2['wdir'],
's1_wspeed': df1['wspeed'],
's2_wspeed': df2['wspeed'],
's1_stn': df1['stn'],
's2_stn': df2['stn'],
}
```

```
In [100]: df = pd.DataFrame(data)

def str2datetime(string):
return datetime.datetime.strptime(string, '%b %d, %Y %I%p')

if use_predict:
headache = [
str2datetime('Sep 5, 2019 7AM'), str2datetime('Sep 5, 2019 10AM'), 8],
[str2datetime('Sep 15, 2019 9AM'), str2datetime('Sep 15, 2019 3PM'), 7],
[str2datetime('Sep 27, 2019 3PM'), str2datetime('Sep 27, 2019 9PM'), 8],
[str2datetime('Oct 15, 2019 2PM'), str2datetime('OCT 15, 2019 6PM'), 7],
[str2datetime('Oct 16, 2019 4AM'), str2datetime('Oct 16, 2019 6PM'), 10],
[str2datetime('Oct 28, 2019 3PM'), str2datetime('OCT 28, 2019 9PM'), 6],
[str2datetime('Nov 28, 2019 5PM'), str2datetime('Nov 28, 2019 6PM'), 6],
[str2datetime('Dec 10, 2019 5AM'), str2datetime('Dec 10, 2019 10AM'), 4],
[str2datetime('Dec 15, 2019 2AM'), str2datetime('Dec 15, 2019 10AM'), 7],
[str2datetime('Dec 20, 2019 1PM'), str2datetime('Dec 20, 2019 6PM'), 9],
]

workout = [
str2datetime('SEP 2, 2019 6PM'), str2datetime('SEP 2, 2019 7PM')],
[str2datetime('SEP 4, 2019 6PM'), str2datetime('SEP 4, 2019 7PM')],
[str2datetime('SEP 6, 2019 6PM'), str2datetime('SEP 6, 2019 7PM')],
[str2datetime('SEP 27, 2019 6PM'), str2datetime('SEP 27, 2019 7PM')],
[str2datetime('OCT 8, 2019 6PM'), str2datetime('OCT 8, 2019 7PM')],
[str2datetime('OCT 20, 2019 8PM'), str2datetime('OCT 20, 2019 9PM')],
[str2datetime('OCT 27, 2019 6AM'), str2datetime('OCT 27, 2019 7AM')],
[str2datetime('OCT 28, 2019 6PM'), str2datetime('OCT 28, 2019 7PM')],
[str2datetime('NOV 2, 2019 6PM'), str2datetime('NOV 2, 2019 7PM')],
[str2datetime('NOV 7, 2019 6PM'), str2datetime('NOV 7, 2019 7PM')],
[str2datetime('NOV 9, 2019 6PM'), str2datetime('NOV 9, 2019 7PM')],
[str2datetime('NOV 12, 2019 6PM'), str2datetime('NOV 12, 2019 7PM')],
[str2datetime('NOV 20, 2019 6PM'), str2datetime('NOV 20, 2019 7PM')],
[str2datetime('DEC 15, 2019 6PM'), str2datetime('DEC 15, 2019 7PM')],
[str2datetime('DEC 16, 2019 6PM'), str2datetime('DEC 16, 2019 7PM')],
[str2datetime('SEP 2, 2020 6PM'), str2datetime('SEP 2, 2020 7PM')],
[str2datetime('SEP 4, 2020 6PM'), str2datetime('SEP 4, 2020 7PM')],
[str2datetime('SEP 6, 2020 6PM'), str2datetime('SEP 6, 2020 7PM')],
]
```

```
In [101]: df['hour'] = pd.Series([0]* df['date'].size)
df['dayweek'] = pd.Series([0]* df['date'].size)
for i in range(df['date'].size):
df['hour'][i] = df['date'][i].hour
df['dayweek'][i] = df['date'][i].weekday()

df['workout'] = pd.Series([0]* df['date'].size)
workout_index = []
for w in workout:
for i in range(df['date'].size):
if w[0] <= df['date'][i] < w[1]:
workout_index.append(i)
df['workout'][workout_index] = 1

df['work'] = pd.Series([0]* df['date'].size)
for i in range(df['date'].size):
if df['date'][i].weekday() == 6 or df['date'][i].weekday() == 0:
continue
if df['date'][i].hour < 9 or df['date'][i].hour > 17:
continue
if str2datetime('OCT 27, 2019 1AM') < df['date'][i] < str2datetime('NOV 3, 2019 11PM'):
continue
df['work'][i] = 1
df['work_hour'] = pd.Series([0] * df['date'].size)
work_sum = 0
for i in range(df['date'].size):
if df['work'][i] == 1:
work_sum += 1
df['work_hour'][i] = work_sum
else:
work_sum = 0

if use_predict:
df['headache'] = pd.Series([0]* df['date'].size)
for w in headache:
for i in range(df['date'].size):
if df['date'][i] >= w[0] and df['date'][i] < w[1]:
df['headache'][i] = w[2]

# del df['date']
if use_predict:
df.to_csv(output_name)
else:
df.to_csv(output_name2020)
```

<ipython-input-101-79c08b35a8fb>:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin g.html#returning-a-view-versus-a-copy
df['hour'][i] = df['date'][i].hour

<ipython-input-101-79c08b35a8fb>:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin g.html#returning-a-view-versus-a-copy
df['dayweek'][i] = df['date'][i].weekday()

<ipython-input-101-79c08b35a8fb>:13: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin g.html#returning-a-view-versus-a-copy
df['workout'][workout_index] = 1

<ipython-input-101-79c08b35a8fb>:23: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin g.html#returning-a-view-versus-a-copy
df['work'][i] = 1

<ipython-input-101-79c08b35a8fb>:29: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin g.html#returning-a-view-versus-a-copy
df['work_hour'][i] = work_sum

<ipython-input-101-79c08b35a8fb>:38: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin g.html#returning-a-view-versus-a-copy
df['headache'][i] = w[2]

```
In [102]: if use_predict and use_average:
data = pd.read_csv('average.csv')
elif use_predict and not use_average:
data = pd.read_csv('normal.csv')
elif not use_predict and use_average:
data = pd.read_csv('2020average.csv')
else:
data = pd.read_csv('2020normal.csv')

df = data.dropna(axis = 0 , how = 'any')

checkFor_nan = df.isnull().values.any()
print (checkFor_nan)
```

False

```
In [103]: if use_predict and use_average:
df.to_csv('DataSetAverage.csv')
elif use_predict and not use_average:
df.to_csv('DataSetNormal.csv')
elif not use_predict and use_average:
df.to_csv('PredictionDataAverage.csv')
else:
df.to_csv('PredictionDataNormal.csv')
```

```
In [104]: import matplotlib.pyplot as plt

plt.matshow(df.corr())
plt.show()
```



