Lab 4.3 — AEP Introducing Holidays

Objective

The objective of this lab is to analyze the Annual Energy Production (AEP) data, introduce holiday effects into the dataset, and examine how holidays influence the energy production values (AEP_MW). This includes identifying missing values, outliers, and performing time-based interpolation considering holidays.

```
import pandas as pd
import numpy as np

df=pd.read_csv(r'C:\Users\PMLS\labreports\
LAB4\3_Outlier_Identified.csv',parse_dates=['Datetime'])
df.head()
```

```
             Datetime   AEP_MW
0 2004-10-01 01:00:00  12379.0
1 2004-10-01 02:00:00  11935.0
2 2004-10-01 03:00:00  11692.0
3 2004-10-01 04:00:00  11597.0
4 2004-10-01 05:00:00  11681.0
```

```
df['Date'] = df['Datetime'].dt.normalize()

df.iloc[0:50]
```

```
              Datetime   AEP_MW       Date
0  2004-10-01 01:00:00  12379.0 2004-10-01
1  2004-10-01 02:00:00  11935.0 2004-10-01
2  2004-10-01 03:00:00  11692.0 2004-10-01
3  2004-10-01 04:00:00  11597.0 2004-10-01
4  2004-10-01 05:00:00  11681.0 2004-10-01
5  2004-10-01 06:00:00  12280.0 2004-10-01
6  2004-10-01 07:00:00  13692.0 2004-10-01
7  2004-10-01 08:00:00  14618.0 2004-10-01
8  2004-10-01 09:00:00  14903.0 2004-10-01
9  2004-10-01 10:00:00  15118.0 2004-10-01
10 2004-10-01 11:00:00  15242.0 2004-10-01
11 2004-10-01 12:00:00  15375.0 2004-10-01
12 2004-10-01 13:00:00  15404.0 2004-10-01
13 2004-10-01 14:00:00  15655.0 2004-10-01
14 2004-10-01 15:00:00  15739.0 2004-10-01
15 2004-10-01 16:00:00  15739.0 2004-10-01
16 2004-10-01 17:00:00  15644.0 2004-10-01
17 2004-10-01 18:00:00  15353.0 2004-10-01
18 2004-10-01 19:00:00  15034.0 2004-10-01
```

```
19 2004-10-01 20:00:00  15211.0 2004-10-01
20 2004-10-01 21:00:00  15349.0 2004-10-01
21 2004-10-01 22:00:00  14837.0 2004-10-01
22 2004-10-01 23:00:00  14067.0 2004-10-01
23 2004-10-02 00:00:00  13147.0 2004-10-02
24 2004-10-02 01:00:00  12260.0 2004-10-02
25 2004-10-02 02:00:00  11672.0 2004-10-02
26 2004-10-02 03:00:00  11352.0 2004-10-02
27 2004-10-02 04:00:00  11177.0 2004-10-02
28 2004-10-02 05:00:00  11142.0 2004-10-02
29 2004-10-02 06:00:00  11331.0 2004-10-02
30 2004-10-02 07:00:00  11866.0 2004-10-02
31 2004-10-02 08:00:00  12387.0 2004-10-02
32 2004-10-02 09:00:00  13144.0 2004-10-02
33 2004-10-02 10:00:00  13712.0 2004-10-02
34 2004-10-02 11:00:00  14082.0 2004-10-02
35 2004-10-02 12:00:00  14080.0 2004-10-02
36 2004-10-02 13:00:00  14056.0 2004-10-02
37 2004-10-02 14:00:00  13934.0 2004-10-02
38 2004-10-02 15:00:00  13758.0 2004-10-02
39 2004-10-02 16:00:00  13579.0 2004-10-02
40 2004-10-02 17:00:00  13620.0 2004-10-02
41 2004-10-02 18:00:00  13483.0 2004-10-02
42 2004-10-02 19:00:00  13379.0 2004-10-02
43 2004-10-02 20:00:00  13825.0 2004-10-02
44 2004-10-02 21:00:00  14056.0 2004-10-02
45 2004-10-02 22:00:00  14015.0 2004-10-02
46 2004-10-02 23:00:00  12940.0 2004-10-02
47 2004-10-03 00:00:00  12172.0 2004-10-03
48 2004-10-03 01:00:00  11443.0 2004-10-03
49 2004-10-03 02:00:00  10807.0 2004-10-03
```

```
len(df)
```

```
121296
```

```
holiday=pd.read_csv(r'C:\Users\PMLS\ML\LAB4\
processed_holiday.csv',parse_dates=['Date'])
```

```
holiday.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 279 entries, 0 to 278
Data columns (total 6 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   Date     279 non-null    datetime64[ns]
 1   Holiday  279 non-null    object
 2   WeekDay  279 non-null    object
 3   Month    279 non-null    int64
```

```
 4   Day      279 non-null    int64
 5   Year     279 non-null    int64
dtypes: datetime64[ns](1), int64(3), object(2)
memory usage: 13.2+ KB
```

```
holiday.head()
```

	Date	Holiday	WeekDay	Month	Day	Year
0	2004-01-01	New Year's Day	Thursday	1	1	2004
1	2004-01-19	Martin Luther King, Jr. Day	Monday	1	19	2004
2	2004-02-14	Valentine’s Day	Saturday	2	14	2004
3	2004-02-16	Washington's Birthday	Monday	2	16	2004
4	2004-04-11	Eastern Easter	Sunday	4	11	2004

```
holiday.drop(['WeekDay','Month','Day','Year'],axis=1, inplace=True)
```

```
(len(holiday)*24)-(21*24) #in 2004 first 9 month is not include which
includes 11 holidays.
                          #similarly in 2018 last 5 month not
included which incluede 10 holidays.  21 total
```

```
6192
```

```
mergedf = pd.merge(df, holiday, on= 'Date', how="left")
mergedf
```

	Datetime	AEP_MW	Date	Holiday
0	2004-10-01 01:00:00	12379.0	2004-10-01	NaN
1	2004-10-01 02:00:00	11935.0	2004-10-01	NaN
2	2004-10-01 03:00:00	11692.0	2004-10-01	NaN
3	2004-10-01 04:00:00	11597.0	2004-10-01	NaN
4	2004-10-01 05:00:00	11681.0	2004-10-01	NaN
...	...	...	...	...
121291	2018-08-02 20:00:00	17673.0	2018-08-02	NaN
121292	2018-08-02 21:00:00	17303.0	2018-08-02	NaN
121293	2018-08-02 22:00:00	17001.0	2018-08-02	NaN
121294	2018-08-02 23:00:00	15964.0	2018-08-02	NaN
121295	2018-08-03 00:00:00	14809.0	2018-08-03	NaN

```
[121296 rows x 4 columns]
```

```
#
mergedf['Holiday'].isna().sum()
```

```
np.int64(115104)
```

```
#121416
121296-mergedf['Holiday'].isna().sum()
```

```
np.int64(6192)
```

```
mergedf['Holiday'] = mergedf['Holiday'].notnull().astype("int")
```

```
mergedf.head()
```

	Datetime	AEP_MW	Date	Holiday
0	2004-10-01 01:00:00	12379.0	2004-10-01	0
1	2004-10-01 02:00:00	11935.0	2004-10-01	0
2	2004-10-01 03:00:00	11692.0	2004-10-01	0
3	2004-10-01 04:00:00	11597.0	2004-10-01	0
4	2004-10-01 05:00:00	11681.0	2004-10-01	0

```
mergedf.to_csv(r'C:\Users\PMLS\ML\
LAB4\4_AEP_Introducing_holidays.csv',index=False)
```

```
mergedf[2000:2060]
```

	Datetime	AEP_MW	Date	Holiday
2000	2004-12-23 09:00:00	16231.0	2004-12-23	0
2001	2004-12-23 10:00:00	16576.0	2004-12-23	0
2002	2004-12-23 11:00:00	16912.0	2004-12-23	0
2003	2004-12-23 12:00:00	16991.0	2004-12-23	0
2004	2004-12-23 13:00:00	16793.0	2004-12-23	0
2005	2004-12-23 14:00:00	16764.0	2004-12-23	0
2006	2004-12-23 15:00:00	16643.0	2004-12-23	0
2007	2004-12-23 16:00:00	16666.0	2004-12-23	0
2008	2004-12-23 17:00:00	16869.0	2004-12-23	0
2009	2004-12-23 18:00:00	17916.0	2004-12-23	0
2010	2004-12-23 19:00:00	18424.0	2004-12-23	0
2011	2004-12-23 20:00:00	18343.0	2004-12-23	0
2012	2004-12-23 21:00:00	18340.0	2004-12-23	0
2013	2004-12-23 22:00:00	18011.0	2004-12-23	0
2014	2004-12-23 23:00:00	17340.0	2004-12-23	0
2015	2004-12-24 00:00:00	16383.0	2004-12-24	1
2016	2004-12-24 01:00:00	15645.0	2004-12-24	1
2017	2004-12-24 02:00:00	15265.0	2004-12-24	1
2018	2004-12-24 03:00:00	15138.0	2004-12-24	1
2019	2004-12-24 04:00:00	15068.0	2004-12-24	1
2020	2004-12-24 05:00:00	15122.0	2004-12-24	1
2021	2004-12-24 06:00:00	15441.0	2004-12-24	1
2022	2004-12-24 07:00:00	15967.0	2004-12-24	1
2023	2004-12-24 08:00:00	16628.0	2004-12-24	1
2024	2004-12-24 09:00:00	17122.0	2004-12-24	1
2025	2004-12-24 10:00:00	17621.0	2004-12-24	1
2026	2004-12-24 11:00:00	17501.0	2004-12-24	1
2027	2004-12-24 12:00:00	17122.0	2004-12-24	1
2028	2004-12-24 13:00:00	16600.0	2004-12-24	1
2029	2004-12-24 14:00:00	16206.0	2004-12-24	1
2030	2004-12-24 15:00:00	15844.0	2004-12-24	1
2031	2004-12-24 16:00:00	15762.0	2004-12-24	1
2032	2004-12-24 17:00:00	16092.0	2004-12-24	1
2033	2004-12-24 18:00:00	17048.0	2004-12-24	1
2034	2004-12-24 19:00:00	17456.0	2004-12-24	1

```
2035 2004-12-24 20:00:00  17407.0 2004-12-24         1
2036 2004-12-24 21:00:00  17474.0 2004-12-24         1
2037 2004-12-24 22:00:00  17772.0 2004-12-24         1
2038 2004-12-24 23:00:00  17618.0 2004-12-24         1
2039 2004-12-25 00:00:00  17147.0 2004-12-25         1
2040 2004-12-25 01:00:00  16669.0 2004-12-25         1
2041 2004-12-25 02:00:00  16218.0 2004-12-25         1
2042 2004-12-25 03:00:00  16135.0 2004-12-25         1
2043 2004-12-25 04:00:00  16107.0 2004-12-25         1
2044 2004-12-25 05:00:00  16229.0 2004-12-25         1
2045 2004-12-25 06:00:00  16470.0 2004-12-25         1
2046 2004-12-25 07:00:00  16935.0 2004-12-25         1
2047 2004-12-25 08:00:00  17548.0 2004-12-25         1
2048 2004-12-25 09:00:00  17927.0 2004-12-25         1
2049 2004-12-25 10:00:00  17837.0 2004-12-25         1
2050 2004-12-25 11:00:00  17453.0 2004-12-25         1
2051 2004-12-25 12:00:00  16891.0 2004-12-25         1
2052 2004-12-25 13:00:00  15967.0 2004-12-25         1
2053 2004-12-25 14:00:00  15088.0 2004-12-25         1
2054 2004-12-25 15:00:00  14564.0 2004-12-25         1
2055 2004-12-25 16:00:00  14394.0 2004-12-25         1
2056 2004-12-25 17:00:00  14745.0 2004-12-25         1
2057 2004-12-25 18:00:00  15856.0 2004-12-25         1
2058 2004-12-25 19:00:00  16502.0 2004-12-25         1
2059 2004-12-25 20:00:00  16678.0 2004-12-25         1
```

```
df.isnull().sum()
```

```
Datetime    0
AEP_MW      0
Date        0
dtype: int64
```