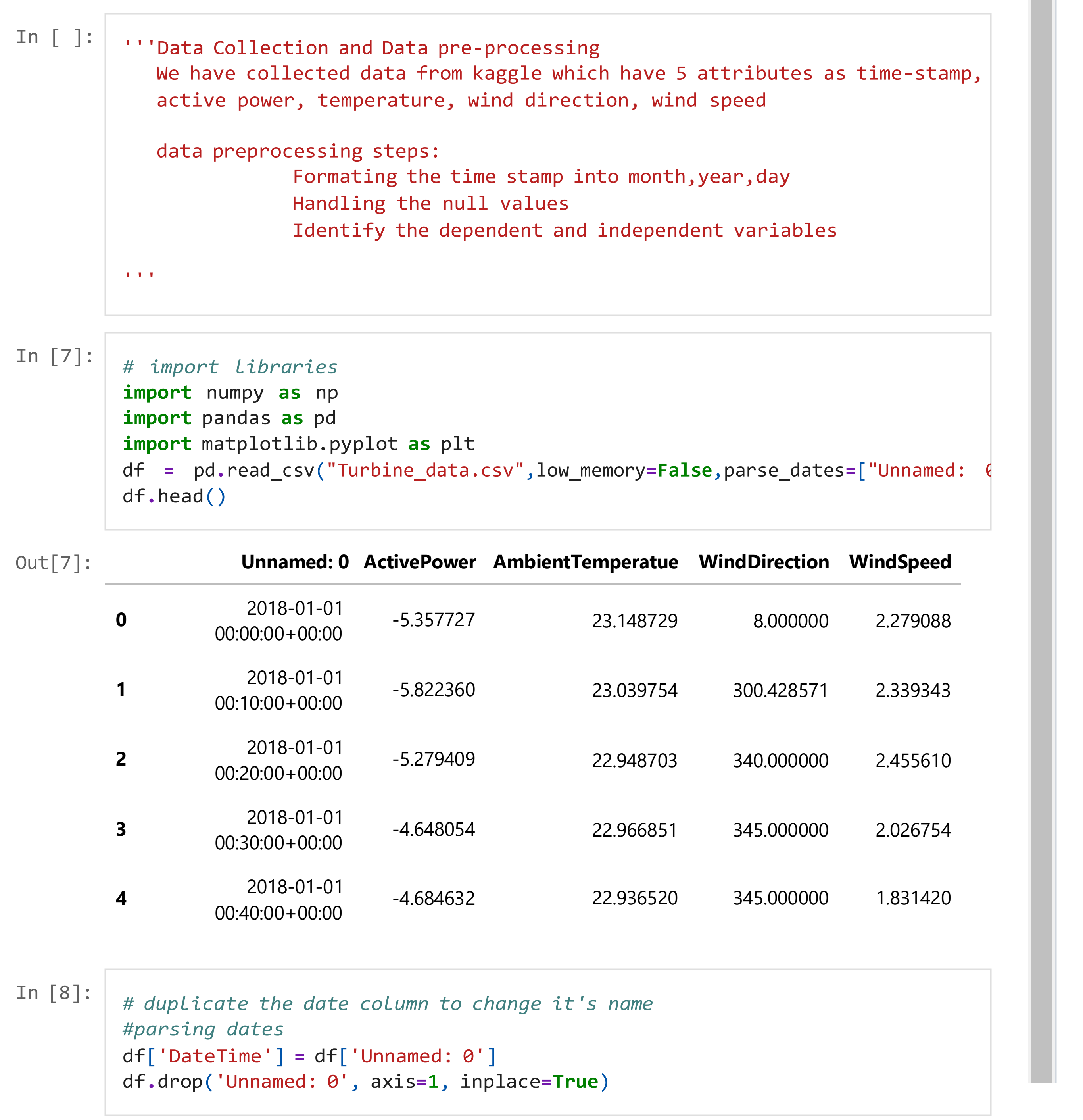
SPRINT 1 :

Date	15 November 2022
Team ID	PNT2022TMID51231
Project Name	Predicting the energy output of wind turbine based on weather condition



In [ ]:

```
'''Data Collection and Data pre-processing
   We have collected data from kaggle which have 5 attributes as time-stamp,
   active power, temperature, wind direction, wind speed

   data preprocessing steps:
               Formating the time stamp into month,year,day
               Handling the null values
               Identify the dependent and independent variables

'''
```

In [7]:

```
# import libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
df  =  pd.read_csv("Turbine_data.csv",low_memory=False,parse_dates=["Unnamed:  0
df.head()
```

Out[7]:

```
           Unnamed: 0  ActivePower  AmbientTemperatue  WindDirection  WindSpeed
0          2018-01-01    -5.357727          23.148729       8.000000   2.279088
       00:00:00+00:00
1          2018-01-01    -5.822360          23.039754     300.428571   2.339343
       00:10:00+00:00
2          2018-01-01    -5.279409          22.948703     340.000000   2.455610
       00:20:00+00:00
3          2018-01-01    -4.648054          22.966851     345.000000   2.026754
       00:30:00+00:00
4          2018-01-01    -4.684632          22.936520     345.000000   1.831420
       00:40:00+00:00
```

In [8]:

```
# duplicate the date column to change it's name
#parsing dates
df['DateTime'] = df['Unnamed: 0']
df.drop('Unnamed: 0', axis=1, inplace=True)
```

In [9]:

```
df['DateTime'].head(20)
```

Out[9]:

```
0     2018-01-01 00:00:00+00:00
1     2018-01-01 00:10:00+00:00
2     2018-01-01 00:20:00+00:00
3     2018-01-01 00:30:00+00:00
4     2018-01-01 00:40:00+00:00
5     2018-01-01 00:50:00+00:00
6     2018-01-01 01:00:00+00:00
7     2018-01-01 01:10:00+00:00
8     2018-01-01 01:20:00+00:00
9     2018-01-01 01:30:00+00:00
10    2018-01-01 01:40:00+00:00
11    2018-01-01 01:50:00+00:00
```

```
12   2018-01-01  02:00:00+00:00
13   2018-01-01  02:10:00+00:00
14   2018-01-01  02:20:00+00:00
15   2018-01-01  02:30:00+00:00
16   2018-01-01  02:40:00+00:00
17   2018-01-01  02:50:00+00:00
18   2018-01-01  03:00:00+00:00
19   2018-01-01  03:10:00+00:00
Name: DateTime, dtype: datetime64[ns, UTC]
```

In [10]:

```
# Add datetime parameters
df['DateTime']  =  pd.to_datetime(df['DateTime'],
 format = '%Y-%m-%dT%H:%M:%SZ',
 errors = 'coerce')

df['year'] = df['DateTime'].dt.year
df['month']  =  df['DateTime'].dt.month
df['day'] = df['DateTime'].dt.day
df['hour'] = df['DateTime'].dt.hour
df['minute']  =  df['DateTime'].dt.minute
```

In [11]:

```
#check for null values
df.isna().sum()
```

Out[11]:

```
ActivePower          23330
AmbientTemperatue    24263
WindDirection        45802
WindSpeed            23485
DateTime                 0
year                     0
month                    0
day                      0
hour                     0
minute                   0
dtype: int64
```

In [14]:

```
#handling null values
df['AmbientTemperatue'].fillna(int(df['AmbientTemperatue'].mean()), inplace=T
df['WindDirection'].fillna(int(df['WindDirection'].mean()), inplace=True)
df['WindSpeed'].fillna(int(df['WindSpeed'].mean()),  inplace=True)
df['ActivePower'].fillna(int(df['ActivePower'].mean()),  inplace=True)
```

In [15]:

```
df.isnull().any()
```

Out[15]:

```
ActivePower          False
AmbientTemperatue    False
WindDirection        False
WindSpeed            False
DateTime             False
year                 False
month                False
day                  False
hour                 False
minute               False
dtype: bool
```

In [16]:

```
#splitting dependent and independent features
independent_features = df[['month','day','AmbientTemperatue','WindDirection',
independent_features.head()
```

Out[16]:

	month	day	AmbientTemperatue	WindDirection	WindSpeed
0	1	1	23.148729	8.000000	2.279088
1	1	1	23.039754	300.428571	2.339343
2	1	1	22.948703	340.000000	2.455610
3	1	1	22.966851	345.000000	2.026754
4	1	1	22.936520	345.000000	1.831420

In [17]:

```
independent_features.isnull().any()
```

Out[17]:

```
month                    False
day                      False
AmbientTemperatue        False
WindDirection            False
WindSpeed                False
dtype: bool
```

In [18]:

```
target = df['ActivePower']
```

In [19]:

```
df_new  =  independent_features
X=np.asanyarray(df_new).astype('int')
y=np.asanyarray(target).astype('int')
print(X.shape)
print(y.shape)
```

```
(118080, 5)
(118080,)
```

In [20]:

```
target.isnull().any()
```

Out[20]:

```
False
```