

Most questions don't need codes to be answered, but I will still use codes to accumulate a complete solution through questions and to provide solid answers instead of just words. Also I'll use codes to analize and calculate statistics that help me understand and answer better. Because I used too much code my answers might seem very long. but this is no true. If you ignored the codes, you'll save time.

1Understanding the Data

Q 1.1 Describe the structure of the data:

```
import pandas as pd

# I personally prefer reading the dataset into a pandas.DataFrame
# object to analyze it and preprocess it easily
df = pd.read_excel("Telemetry_Dumps_ssbums_18-11-2025_ZN.10-
PB.1_child__1_ch3pm.xlsx")

print('Feature names: ', sorted(df.columns)) # I sorted the columns to
gather similar quantities together, g.e. 'Amp1', 'Amp2', 'Amp3'.
print('Features count: ', len(df.columns))

print('samples count: ', len(df))

Feature names:  ['Amp1', 'Amp2', 'Amp3', 'Hz', 'PF1', 'PF2', 'PF3',
'PFA', 'TkVAr', 'TkW', 'V1', 'V2', 'V3', 'Vn1', 'Vn2', 'Vn3',
'datetime', 'kVAr1', 'kVAr2', 'kVAr3', 'kVArh', 'kW1', 'kW2', 'kW3',
'kWh', 'timestamp']
Features count:  26
samples count:  8882
```

As it's clear above, this is a tabular dataset with 26 features/columns and 8882 samples/rows

```
# display the first five samples/rows
df.head()

      timestamp    PF1     kW1   kVAr1      kWh    Amp1   kVArh      Hz     Vn1
Amp2 \
0  1759266061998  0.6  0.308   -5.6  7123.6    2.4    8.33  59.94  227.8
4.4
1  1759266363510  0.6  0.308   -5.6  7124.0    2.4    8.33  59.99  228.1
4.4
2  1759266664943  0.6  0.312   -5.6  7124.0    2.4    8.33  60.00  228.4
4.4
3  1759266966433  0.6  0.308   -5.6  7124.0    2.4    8.33  59.98  228.5
4.4
4  1759267267961  0.6  0.308   -5.6  7124.0    2.4    8.33  60.00  228.4
4.4
```

	...	kVAr3	TkVAr	kW2	TkW	PFa	kVAr2	PF3	PF2	kW3	\
0	...	6.1	-19.3	-0.040	0.832	0.39	-19.8	0.92	-0.04	0.564	
1	...	6.1	-19.3	-0.040	0.832	0.39	-19.8	0.92	-0.04	0.564	
2	...	6.0	-19.4	-0.036	0.840	0.39	-19.8	0.92	-0.04	0.564	
3	...	6.0	-19.5	-0.036	0.836	0.39	-19.9	0.92	-0.04	0.564	
4	...	8.2	-17.1	-0.044	0.868	0.39	-19.7	0.88	-0.04	0.604	

	datetime											
0	2025-10-01	12-01-01	AM									
1	2025-10-01	12-06-03	AM									
2	2025-10-01	12-11-04	AM									
3	2025-10-01	12-16-06	AM									
4	2025-10-01	12-21-07	AM									

[5 rows x 26 columns]

Q 1.2 Explain what each column represents:

First, I have to mention that in this proposed building power system, we are using three phases. Each phase is an independent AC wave with a 120 degrees shift from the others. This explain why sometimes there are three different values/features for the same quantity (e.g. 'PF1', 'PF2', 'PF3'). Knowing that, I'll explain each quantity only once, and I won't repeat it for every phase.

- **Amp:** The electrical current that is currently being drawn at a specific phase.
- **Hz:** The frequency, g.e. 50Hz or 60Hz.
- **PF:** This is the power factor which measures the electrical power efficiency at a specific phase.
- **PFa:** I searched a lot and tried several equations trying to calculate this myself in order to figure out what is it but I didn't find an answer.
- **TkVAr:** Total reactive power (i.e. kVAr1 + kVAr2 + kVAr3). I verified this using `df[['kVAr1', 'kVAr2', 'kVAr3']].sum(axis=1)`
- **TkW:** Total active power (i.e. kW1 + kW2 + kW3). I verified this using `df[['kW1', 'kW2', 'kW3']].sum(axis=1)`
- **V:** The Voltage at a specific phase.
- **Vn:** The Voltage difference between a specific phase Voltage and the neutral cable Voltage.
- **datetime:** The date at which the measurement was taken.
- **kVAr:** This is the reactive power at a specific phase, which is a part of power that doesn't get consumed in real work.
- **kVArh:** This is the accumulated reactive energy over time.
- **kW:** This is the active power component at a specific phase that does work.
- **kWh:** This is the accumulated active energy over time.
- **timestamp:** The time at which the measurement was taken.

Q 1.3 Identify potential issues in the data (missing values, noise, anomalies, duplicates, misalignment, etc.):

- missing values:

By inspecting the data file by eyes, I guess there are no missing values. But this method isn't reliable enough. Thus, the following code checks whether there is any unassigned value, and count these values at each column:

```
df.isna().sum()  
timestamp      0  
PF1            0  
kW1            0  
kVAr1          0  
kWh             0  
Amp1           0  
kVArh          0  
Hz              0  
Vn1             0  
Amp2           0  
V2              0  
Amp3           0  
V1              0  
Vn2             0  
V3              0  
Vn3             0  
kVAr3          0  
TkVAr          0  
kW2             0  
TkW             0  
Pfa             0  
kVAr2          0  
PF3             0  
PF2             0  
kW3             0  
datetime        0  
dtype: int64
```

These zeros indicate that all the cells in the dataset are assigned to a value. But, while it is true that this is fine for the 25 numerical data type columns we have, the "datetime" column is string, I might not have a true value if it is assigned to empty string (" or ""). Thus, I use the following code to check for this:

```
(df['datetime'] == '').sum()  
np.int64(0)
```

- **duplicates:**

If I'm not wrong, it is not an issue if any of the readings got repeated/suplicated. This is an issue only for "datetime" and "timestamp".

```
print('total number of duplicated "datetime" values: ',  
df['datetime'].duplicated().sum())  
print('total number of duplicated "timestamp" values: ',  
df['timestamp'].duplicated().sum())  
  
total number of duplicated "datetime" values: 0  
total number of duplicated "timestamp" values: 0
```

- **anomalies & outliers:**

To detect and check if we have any anomalies, I'll use Interquartile range rule (IQR). Since I studied this statistical method before, and it's simple and fast to implement.

```
for column in sorted(df.columns.drop('datetime')):  
    Q1 = df[column].quantile(0.25)  
    Q3 = df[column].quantile(0.75)  
    IQR = Q3 - Q1  
  
    anomalies = df[(df[column] < Q1 - 1.5 * IQR) | (df[column] > Q3 +  
1.5 * IQR)][column]  
    print('feature/column: ', column)  
    print('lower limit: ', Q1 - 1.5 * IQR)  
    print('upper limit: ', Q3 + 1.5 * IQR)  
    print('avg: ', df[column].mean())  
    print('std: ', df[column].std())  
    print('min: ', df[column].min())  
    print('max: ', df[column].max())  
    print('anomalies count: ', len(anomalies))  
    print('unique anomalies: ', anomalies.unique(), '\n')  
  
feature/column: Amp1  
lower limit: -0.20000000000000018  
upper limit: 4.6000000000000005  
avg: 2.246656158522855  
std: 0.6883156021705701  
min: 1.2  
max: 8.8  
anomalies count: 20  
unique anomalies: [5.2 4.8 8.8]  
  
feature/column: Amp2  
lower limit: 3.3999999999999995  
upper limit: 5.0000000000000001  
avg: 4.744562035577572  
std: 1.8372486916326265
```

```
min: 2.8000000000000003
max: 18.4
anomalies count: 1599
unique anomalies: [11.6 8.8 17.2 16.4 3.2 10.4 10.8 11.2 8. 9.2
5.2 12.8 12. 7.2
7.6 8.4 15.6 17.6 12.4 9.6 14. 16.8 6. 5.6 10. 13.6 14.4 18.4
16. 2.8 6.4 14.8 18. 6.8]

feature/column: Amp3
lower limit: 1.4
upper limit: 6.200000000000001
avg: 4.015537041206936
std: 1.9500234838250388
min: 2.800000000000003
max: 23.6
anomalies count: 500
unique anomalies: [15.6 9.2 8.8 10.4 10.8 16. 16.4 8.4 20.8 14.4
10. 16.8 17.2 9.6
15.2 11.2 6.4 17.6 23.2 23.6 12.8 11.6 21.2 12. 6.8 22.8 12.4
7.6]

feature/column: Hz
lower limit: 59.87999999999995
upper limit: 60.12000000000005
avg: 60.000275838775046
std: 0.03330772911373938
min: 59.91
max: 60.08
anomalies count: 0
unique anomalies: []

feature/column: PF1
lower limit: 0.4850000000000015
upper limit: 0.6849999999999998
avg: 0.5934744426930871
std: 0.04157248422318849
min: 0.52
max: 0.97
anomalies count: 226
unique anomalies: [0.69 0.7 0.71 0.72 0.73 0.74 0.75 0.97]

feature/column: PF2
lower limit: -0.1
upper limit: 0.06
avg: 0.005951362305786985
std: 0.11519663013798953
min: -0.21
max: 0.46
anomalies count: 1674
unique anomalies: [ 0.34 0.27 0.28 0.22 0.45 0.39 0.23 -0.16 -
```

```
0.15 -0.14 -0.12 -0.13
 0.35 -0.11  0.21  0.3   0.29  0.24  0.26  0.42 -0.17  0.33  0.2
0.4
 0.43  0.25  0.41  0.37  0.36  0.31  0.32  0.08  0.07  0.44  0.17
0.46
 0.38  0.19 -0.2  -0.21 -0.18  0.1   0.18  0.11  0.12  0.16  0.13
0.14]

feature/column:  PF3
lower limit:  0.6999999999999998
upper limit:  1.02
avg:  0.8502893492456655
std:  0.06203122261259279
min:  0.38
max:  0.94
anomalies count:  291
unique anomalies:  [0.61 0.67 0.65 0.64 0.66 0.57 0.6  0.68 0.62 0.49
0.39 0.59 0.69 0.63
 0.38 0.58]

feature/column:  PFa
lower limit:  0.3100000000000005
upper limit:  0.55
avg:  0.43454852510695785
std:  0.04644220799251726
min:  0.34
max:  0.69
anomalies count:  287
unique anomalies:  [0.56 0.57 0.63 0.59 0.6  0.64 0.58 0.61 0.62 0.69]

feature/column:  TkVAr
lower limit:  -25.84999999999998
upper limit:  -5.45
avg:  -16.406113487953164
std:  12.374881473388719
min:  -87.6
max:  73.4
anomalies count:  1122
unique anomalies:  [-57.9 -43.2  19.  -58.2  45.6  12.1 -81.2 -83.3 -
56.4   8.4   8.3   8.2
 8.1   8.5  -50.7 -52.5 -52.2 -54.6  19.9  20.3 -40.8 -39.5 -57.
44.
 17.8 -56.  -42.7 -55.3 -56.8 -42.6   4.5 -41.9 -58.9 -59.9 -0.5 -
59.2
 43.6 -56.3 -32.4 -57.4 -41.7 -54.1 -54.5   9.3   9.  9.6   8.
6.6
 6.7   5.2   6.   5.9   6.3   5.5   5.7   6.1   6.5   8.7   10.2
10.1
 10.5  10.3  10.4  9.9   7.8   7.5   7.7   7.3   7.1   7.4   7.2
```



```
-40.6 -36. -36.2 -38.4 -36.4 -53.4 6.4 -34.8 6.2 41.3 31.1  
73.4  
8.6 10.7 11. -31.2 -30.9 11.1 -31.1 40.5 4.7 5.6 4.3 -  
1.9  
40.7 51. 20.6 25.3 19.5 -42.4 -52.9 18.7 -40.5 -56.7 -39.3 -  
73.4  
-61.7 46.8 -28.1 -59.4 -37.6 42.2 20.2 1.8 47.9 -40.1 -78.9  
49.6  
-53.6 41.9 21.5 44.8 -75.7 5.1 -37.7 -51.4 14.6 -83.1 -58.1  
19.3  
45.2 -57.6 43.9 45.9 -81.6 -38. 14.4 17.6 27.3 66.9 -35.2  
51.5  
53.9 -49.8 -54.9 -31. 21.7 18.1 -3.5 23.3 16.1 -60.7 -43.8  
4.8  
43.8 45.3 28.5 -49.2 -48.5 19.7 -64.9 -66.9 2.6 24.8 -63.1 -  
47.9  
-48.7 -81.8 -39.8 -50.3 -49.5 -48.4 -4.2 -4.3 -34.5 -27.1 -3.9 -  
36.6  
-74.7 -27.7 -2.9 -75.4 -48.9 -4.8 -34.3 -46.5 -35.6 -33.7 -44. -  
3.  
-2.3 -2.7 -32.9 -86. -87.6 -85.4 -57.5 -52.7 -5.4 -26.5 -26.7 -  
27.3  
-28.4 -29.7 -29.5 -27.9 -27.2 -33.2 -27.6 -29.8 -30.2 -27.4 -26.4 -  
29.6  
-29.1 -27.8 -26.6]
```

```
feature/column: TkW  
lower limit: 0.4539999999999998  
upper limit: 1.5260000000000002  
avg: 1.097741049313218  
std: 0.43403289044105997  
min: 0.704  
max: 5.004  
anomalies count: 1010  
unique anomalies: [1.84 3.024 1.868 2.428 1.936 4.22 2.492 2.756  
1.896 1.812 1.816 1.8  
1.832 1.836 1.828 1.864 1.964 1.952 1.972 2.872 2.244 2.188 2.216  
1.78  
1.668 2.06 2.696 1.908 1.788 2.204 2.04 1.66 3.596 1.744 2.056  
2.112  
3.868 2.172 2.704 2.076 1.552 1.86 2.028 1.808 1.892 1.88 1.912  
1.872  
1.9 1.884 1.916 1.844 1.824 1.796 1.792 1.82 1.856 1.888 1.876  
1.772  
1.776 1.768 1.764 2.408 1.804 1.76 2.348 1.756 2.34 1.752 2.712  
2.708  
3.876 3.112 2.044 3.544 2.608 3.536 3.28 2.052 2.956 2.036 2.672  
2.08  
3.676 3.068 2.472 2.088 2.852 2.3 2.868 2.048 2.236 1.924 1.984
```

```

1.712
2.152 1.664 1.54 1.548 1.56 1.556 2.312 1.528 1.572 1.536 1.532
3.052
2.296 2.14 3.296 1.996 2.032 2.164 2.724 2.332 2.148 2.484 2.464
1.576
1.564 2.452 2.436 2.16 2.136 2.224 1.728 3.684 1.944 2.108 1.692
1.688
3.42 2.836 1.976 2.808 2.092 2.352 2.072 1.724 2.288 1.74 1.968 2.7
2.28 2.6 1.716 2.572 2.44 2.26 1.736 3.572 3.836 2.376 3.744
2.82
1.684 2.316 2.124 1.96 3.232 2.384 2.848 2.084 2.104 2.252 2.528
1.704
2.656 3.176 2.168 2.268 2.292 2.536 1.58 1.568 3.792 2.884 1.928
2.212
3.644 1.624 1.584 1.616 2.468 2.752 3.64 2.156 2.256 2.248 1.648
2.964
2.932 2.56 1.696 1.94 2.948 1.68 1.612 2.128 2.192 2.184 1.932
2.232
1.748 2.648 2.784 2.788 2.176 2.308 3.724 2.116 1.732 2.392 1.604
2.812
2.284 2.512 4.072 2.448 2.432 2.1 3.088 2.22 2.12 1.6 3.336
2.676
1.62 2.456 1.848 2.064 3.02 2.776 1.92 3.096 2.096 2.816 2.196
1.632
1.644 2.86 2.576 2.016 1.636 1.672 2.024 1.652 1.64 3.524 3.452
3.196
2.32 2.324 2.416 2.068 2.68 3.476 2.132 1.956 1.98 1.676 2.94
4.924
3.356 5.004 2.336 3.1 2.588 2.596 2.344 2.612 2.46 1.992 2.764
2.84
1.544 2.276 3.752 2.444 2.2 4. 3.444 2.008 2.74 2.54 2.728
3.96
2.548 2.144 3.064 4.416 3.228 3.244 4.052 1.708 1.72 2.66 2.624
1.904
2.012 1.988 3.784 3.08 2.792 2. 3.9 2.18 2.38 3.3 2.644
2.804
1.948 2.424 2.412 2.828 2.664 3.624 2.4 2.72 3.74 3.012 1.596
2.592
3.192 2.604 1.852 3.252 2.524 2.264 4.508 3.852 2.388 4.524 3.044
3.316
3.004 2.568 3.164 2.684 1.608 2.52 2.856 2.208 2.372 2.688 1.628
2.904
2.368 2.508]

```

feature/column: V1
lower limit: 390.6500000000001
upper limit: 399.8499999999999
avg: 395.2319635217293
std: 1.567254990693302

```
min: 389.7
max: 399.7
anomalies count: 5
unique anomalies: [390.3 390. 389.7 390.5 390.4]

feature/column: V2
lower limit: 390.1500000000001
upper limit: 399.3499999999999
avg: 394.74079036253096
std: 1.5879150226293235
min: 389.1
max: 399.2
anomalies count: 13
unique anomalies: [389.8 390.1 389.9 390. 389.6 389.5 389.1]

feature/column: V3
lower limit: 390.25
upper limit: 399.45000000000005
avg: 394.83607295654133
std: 1.5541251022105784
min: 389.4
max: 399.5
anomalies count: 9
unique anomalies: [389.8 389.7 399.5 390.2 390.1 390. 389.4]

feature/column: Vn1
lower limit: 225.6499999999998
upper limit: 230.85000000000002
avg: 228.27005179013736
std: 0.8988935745635759
min: 225.1
max: 231.0
anomalies count: 7
unique anomalies: [225.6 225.1 225.5 225.4 231. ]

feature/column: Vn2
lower limit: 225.3999999999998
upper limit: 231.0
avg: 228.16062823688358
std: 0.9470167034901539
min: 224.5
max: 230.6
anomalies count: 13
unique anomalies: [225.1 224.8 225.3 225. 224.5 224.7]

feature/column: Vn3
lower limit: 225.1499999999998
upper limit: 230.35000000000002
avg: 227.7038392253997
std: 0.9348534795940939
```

```
min: 223.4
max: 230.4
anomalies count: 53
unique anomalies: [225.1 224.6 224.7 224.9 225. 223.9 224.8 230.4
224.4 224.2 223.4]

feature/column: kVAr1
lower limit: -7.3500000000000005
upper limit: -3.749999999999999
avg: -5.6515987390227425
std: 0.6437649556198372
min: -8.2
max: -0.5
anomalies count: 166
unique anomalies: [-7.5 -7.4 -7.7 -7.6 -7.9 -7.8 -8. -8.2 -8.1 -0.5]

feature/column: kVAr2
lower limit: -22.65
upper limit: -16.65
avg: -21.94167980184643
std: 10.06369753732004
min: -92.3
max: -13.3
anomalies count: 1976
unique anomalies: [-60.2 -45.6 -45.5 -60.3 -86.4 -85.5 -87.6 -60.7 -
15.1 -15.2 -15.3 -14.9
-15. -15.6 -14.8 -55.1 -55.5 -55.3 -15.7 -56.2 -16.1 -15.5 -57.4 -
42.7
-60. -58.8 -47.4 -60.1 -22.8 -59.7 -45.1 -15.9 -16.3 -60.8 -23. -
46.1
-61.9 -61.6 -22.7 -64.2 -62.6 -23.1 -23.2 -37. -16.4 -16.5 -56.7 -
16.6
-16.2 -16. -15.4 -41.2 -41.1 -40.7 -40.9 -40.3 -40.4 -42. -80.2 -
42.1
-56.5 -43.1 -43.2 -36.8 -86.9 -24.4 -88.9 -63.7 -49.5 -23.4 -23.3 -
58.6
-62.3 -59.5 -43.7 -58.4 -43.8 -57.3 -58.2 -15.8 -56.8 -48.2 -45.3 -
44.9
-44.8 -45. -45.2 -45.4 -44.4 -44.6 -44.5 -42.6 -45.8 -60.4 -46.3 -
62.5
-62. -58.1 -24.7 -61.5 -47.2 -61.8 -25.6 -24.3 -23.9 -23.6 -75.4 -
48.5
-24. -64. -64.3 -64.5 -64.6 -63. -63.2 -58.3 -47.3 -47.7 -43.5 -
60.6
-24.9 -23.5 -46.5 -86.6 -61.7 -48.1 -61.4 -22.9 -63.1 -47.6 -57.7 -
59.6
-46.8 -57.6 -42.3 -59.9 -57.8 -42.9 -56.4 -45.7 -60.9 -61.2 -46. -
61.
-26.8 -25.2 -56.6 -59. -63.3 -59.8 -49.1 -14.6 -14.7 -46.7 -61.1 -
```

```
23.8
 -25.9 -25.4 -56.3 -27.6 -53.7 -25.5 -26. - -63.6 -25.7 -49.4 -52.4 -
89.5
 -26.5 -27.2 -67.8 -51.9 -24.2 -45.9 -56.9 -25.1 -65. - -24.5 -36. -
64.4
 -83.7 -61.3 -57.2 -58.9 -62.2 -60.5 -58.5 -57.9 -59.4 -37.6 -38.6 -
62.1
 -38.5 -88. - -51.2 -29.9 -50.7 -48.9 -59.3 -44.1 -44.7 -35.9 -46.4 -
44.3
 -46.2 -68.2 -47. - -46.6 -41.3 -23.7 -85.6 -84.8 -57.1 -40.6 -65.8 -
92.3
 -26.2 -66.3 -25. - -26.6 -26.7 -55.6 -58. - -46.9 -72. - -62.9 -62.4 -
33.4
 -43.3 -83. - -59.1 -13.7 -13.3 -13.6 -13.5 -14.4 -14.5 -41. - -82.7 -
63.9
 -34.2 -72.8 -66.5 -62.7 -35.7 -42.5 -47.9 -40.8 -40.5 -24.8 -80.7 -
41.8
 -55.9 -81.5 -63.5 -50.1 -47.1 -76. - -38.1 -81.7 -42.8 -87.8 -26.3 -
84.5
 -41.7 -65.3 -26.1 -26.4 -67.1 -52.1 -62.8 -38.2 -65.1 -51.1 -44.2 -
43.9
 -35.4 -59.2 -41.4 -55.2 -71.1 -71.3 -14.1 -42.4 -24.1 -66. - -51.4 -
50.6
 -33.3 -49.6 -83.5 -83.9 -27.5 -57. - -28.1 -42.2 -41.9 -80.9 -90.6 -
91.5
 -49.7 -86.7 -25.3 -24.6 -27.8 -31.7 -31.8 -32.9 -29.5 -35.2 -32.3 -
32.
 -31.9 -34.7 -32.5 -30.8 -31.6 -33.5 -33.9 -30.4 -31. - -32.1]
```

```
feature/column: kVAr3
lower limit: 1.1499999999999995
upper limit: 17.55
avg: 11.189236658410266
std: 9.15757187996305
min: 5.5
max: 100.1
anomalies count: 536
unique anomalies: [ 70.2  70.8  38.8  69.7  29.2  29.1  29.3  29.4
44.3  45.1  44. - 44.1
 72. - 44.4  71.9  70.6  72.4  31.1  30.8  30.6  30.4  31.4  30.3
30.2
 28.1  28. - 28.2  28.4  32.2  31.6  31.3  31.2  30.9  30.7  28.7
28.5
 28.3  31. - 30.1  30.5  90.9  59. - 91.2  91.1  42.6  41.7  40.7
74.6
 77.7  46.8  38.7  38.1  72.8  41.1  73.4  42.4  41.9  41.3  72.2
43.8
 43.2  42.3  43.4  43.6  43.3  40.5  74.3  43.7  44.6  76.3  73.6
76.5
```

```
    77.4  69.1  68.   68.3  37.6  42.5  42.9  43.   73.3  44.5  73.2  
74.  
    71.3  69.9  71.8  69.3  42.8  44.7  73.1  40.9  45.7  45.4  47.6  
27.9  
    39.8  38.2  73.   72.9  43.9  74.7  47.8  76.4  45.8  39.5  39.2  
41.8  
    38.3  38.   45.3  39.4  70.7  70.4  40.   40.1  36.8  70.5  39.3  
74.5  
    40.4  75.1  44.9  74.9  32.3  45.6  46.2  74.4  42.2  41.2  41.5  
42.1  
    74.1  73.5  78.   40.3  36.3  99.4  69.2  37.9  100.1 49.2  35.9  
35.8  
    36.   42.7  76.8  48.3  45.9  44.8  44.2  71.6  47.5  47.4  74.2  
20.9  
    36.1  38.4  62.5  92.7  29.8  93.2  27.8  32.4  31.7  29.9  31.5  
32.8  
    62.8  32.   30.   29.   28.9  63.9  66.2  33.2  32.9  35.   34.5  
33.7  
    43.1  48.4  46.   45.   34.8  68.7  75.5  73.8  72.3  69.4  50.  
71.2  
    41.4  20.1  19.2  39.7  97.7  98.1  35.6  36.2  77.5  18.5  46.5  
68.8  
    42.   43.5  52.6  52.4  53.8  17.8  18.1  48.5  18.3  18.   20.5  
19.1  
    18.6  21.1  22.2  21.4  17.6  49.8  20.6  18.7  18.4  19.5  19.  
18.9  
    17.7  18.2  20.4  19.6  47.2  20.3  51.3]
```

feature/column: kVArh

lower limit: 8.33

upper limit: 8.33

avg: 8.33

std: 0.0

min: 8.33

max: 8.33

anomalies count: 0

unique anomalies: []

feature/column: kW1

lower limit: -0.03399999999999995

upper limit: 0.59

avg: 0.3071195676649403

std: 0.110421409280187

min: 0.16

max: 1.92

anomalies count: 138

unique anomalies: [0.592 0.596 0.62 0.628 0.696 0.604 0.636 0.684

0.78 0.724 0.704 0.72

0.728 0.672 0.768 0.692 0.644 0.664 0.608 0.624 0.632 0.64 0.652

```

0.732
 0.7   0.656  0.66   0.648  0.612  0.6    0.712  0.668  0.736  0.756  0.616
0.708
 0.688  0.74   0.744  1.92  ]

feature/column:  kW2
lower limit:  -0.09
upper limit:  0.05399999999999999
avg:  0.04963746903850484
std:  0.2420518957111307
min:  -0.132
max:  1.832
anomalies count:  1265
unique anomalies:  [ 0.884  0.544  0.552  0.892  0.224  0.228  1.752
1.476  0.24   1.76
 0.9   -0.112 -0.108 -0.104 -0.1    0.828 -0.092  0.836  0.852  0.176
 0.18   0.172  0.592  0.52   0.888  0.876  0.6    0.924  0.912  0.568
 0.232  0.264  0.916  0.268  1.204  0.932  0.444  0.272  0.276  0.564
 0.872  0.832 -0.096 -0.116  0.472  0.468  0.464  0.808  0.804  0.484
 0.48   0.152  0.156  0.74   0.16   1.388  0.164  1.084  0.168  0.492
 1.096  0.512  0.516  0.436  1.744  0.3    1.804  1.208  0.608  0.928
 0.208  1.136  0.784  1.14   0.216  0.8    0.86   0.54   0.856  0.604
 0.556  0.572  0.576  0.58   0.92   1.164  0.824  0.244  0.848  0.236
 0.22   0.06   1.24   0.292  0.288  1.22   1.232  1.228  0.94   0.944
 0.28   0.536  0.96   0.064  0.588  0.584  1.524  0.252  0.26   0.844
 1.196  1.192  0.596  0.88   0.948  1.2    0.256  0.532  0.936  0.528
 1.152  1.156  0.816  0.56   1.124  1.188  1.212  0.676  0.524  1.168
 1.18   0.976  0.068  0.764  0.772  0.108  0.708  0.092  0.088  0.096
 0.084  0.08   0.652  0.98   0.328  0.34   0.336  0.64   0.704  1.572
 0.332  0.348  0.364  1.292  0.076  0.072  1.016  1.    0.248  0.864
 0.2   1.132  1.184  0.192  0.896  0.908  0.452  0.212  0.972  0.488
 1.8   0.304  0.284  0.664  0.868  0.904  0.432  0.84   1.344  0.508
 0.296  0.308  0.312  1.772  0.624  1.504  0.956  0.504  0.992  0.62
 1.832  1.224  0.316  0.184  1.12   0.188  0.82   1.16   1.448  0.796
 0.204  1.148 -0.128 -0.132 -0.124 -0.12   1.428  1.544  1.216  1.244
 0.368  1.092  0.756  1.412  1.108  1.42   0.5    1.128  1.372  0.952
 1.176  1.512  0.548  0.964  0.056  1.416  0.196  0.668  1.56   0.132
 0.66   1.516  1.236  0.628  0.968  1.172  1.464  1.304  0.984  0.696
 1.468  0.688  0.616  1.46   1.5    0.136  1.436  1.644  0.632  1.664
 0.988  0.612  1.536  0.112  0.416  0.32  ]

feature/column:  kW3
lower limit:  0.42
upper limit:  0.964
avg:  0.7399247917135781
std:  0.2721386541956234
min:  0.4
max:  3.4
anomalies count:  504
unique anomalies:  [ 2.16   1.4    1.616  1.612  1.608  1.604  1.6    1.596

```

```
1.592 1.584 1.524 1.544
1.52 1.532 2.224 1.452 2.22 2.192 2.24 1.632 1.624 1.62 1.648
1.556
1.56 1.564 1.548 1.588 1.644 1.636 1.628 1.568 1.572 1.552 1.58
3.064
2.308 3.052 1.464 1.492 1.468 2.372 1.528 1.316 1.392 2.132 1.372
2.128
2.12 1.44 2.264 2.28 1.476 1.508 1.54 1.512 1.536 1.444 2.304
2.336
2.288 2.356 2.08 2.06 2.056 1.336 1.436 1.488 1.504 2.276 2.152
2.116
2.228 1.408 1.38 2.188 2.18 1.668 2.34 1.348 1.268 1.356 2.104
2.096
1.376 1.308 1.404 1.5 1.46 2.312 1.256 2.284 1.48 1.472 2.268
2.344
1.876 3.4 2.688 1.928 2.696 2.676 3.388 1.916 2.684 2.184 1.884
1.88
2.26 1.692 2.328 2.208 2.204 1.64 0.4 1.896 2.4 3.128 3.096
1.576
1.652 1.66 1.688 1.684 1.676 1.68 1.664 1.72 1.712 1.728 2.432
1.696
1.708 1.716 2.492 2.544 1.76 1.748 1.776 1.756 1.752 2.296 2.196
1.344
2.212 1.416 2.324 2.136 2.232 2.272 1.388 2.164 2.2 0.972 2.256
1.448
1.94 1.932 1.892 3.356 1.888 3.352 3.36 1.912 1.872 2.32 2.084
2.088
1.396 0.976 0.98 0.988 0.984 1.172]
```

```
feature/column: kwh
lower limit: 6715.75
upper limit: 8436.95
avg: 7570.754109434812
std: 246.82518726831302
min: 7123.6
max: 7980.0
anomalies count: 0
unique anomalies: []
```

```
feature/column: timestamp
lower limit: 1757926936207.125
upper limit: 1763283352118.125
avg: 1760605148866.8376
std: 773261157.4963155
min: 1759266061998
max: 1761944263643
anomalies count: 0
unique anomalies: []
```

Anomalies discussion:

- **Currents (Amp1, Amp2, Amp3):** These readings are not stable. They have huge std. values, wide IQR limits, and some outliers are very extreme and far from average.
- **Frequency (Hz):** The readings are precise and perfect.
- **Power Factors (PF1, PF2, PF3, PFA):** The readings were mostly precise. Even most of the outliers were fine. But, some outliers were extreme than to be ignored. But such outliers are very few. Thus, I guess we can later replace these few values with the average or the closest limit.
- **Voltage Readings (V1, V2, V3, Vn1, Vn2, Vn3):** These readings are very stable and almost always close to the average. Also none of the outliers were very extreme or weird. Thus, in my humble opinion, I think these outliers can just be ignored.
- **Powers (TkVar, TkW, kVar1, kVar2, kVar3, kW1, kW2, kW3):** These readings are not stable. They have huge std. values, wide IQR limits, and some outliers are very extreme and far from average.
- **Accumulated Energies (kVarh, kWh):** The readings are precise and perfect.

• **anomalies & noise:**

I will define a noise to be a physically wrong value or a weird behaviour of the data. The previous analysis showed that the Frequency, Voltages, and Accumulated Energies readings are fine and don't suffer from any weird behaviour or wrong values. Thus, these readings will be skipped at this step. But I'll just check the rest.

Note: in this definition, anomalies and noise definitions can intersect.

- **physically wrong values:** These are the values that don't make sense, such as the current having a negative value, the scale of the power factor is greater than 1.0, and accumulated energies decrease. It is clear above that the first two issues don't exist. But I am not sure about the last one. I'll use the following code to check for this:

```
print('the number of the active energy decreased: ', (df['KWh'].diff() < 0).sum())
print('the number of the reactive energy decreased: ',
      (df['kVarh'].diff() < 0).sum())

the number of the active energy decreased:  0
the number of the reactive energy decreased:  0
```

- **sudden changes:** Noise usually changes the values fast and instantly. The following code illustrates this behaviour visually:

```
import matplotlib.pyplot as plt

plt.figure(figsize=(18, 8))

plt.subplot(3, 2, 1)
plt.plot(df[['Amp1', 'Amp2', 'Amp3']], label=['Amp1', 'Amp2', 'Amp3'])
plt.title('Currents')
plt.xlabel("sample")
plt.ylabel('value')
```

```
plt.legend()

plt.subplot(3, 2, 2)
plt.plot(df[['PF1', 'PF2', 'PF3', 'PFa']], label=['PF1', 'PF2', 'PF3', 'PFa'])
plt.title('Power Factors')
plt.xlabel("sample")
plt.ylabel('value')
plt.legend()

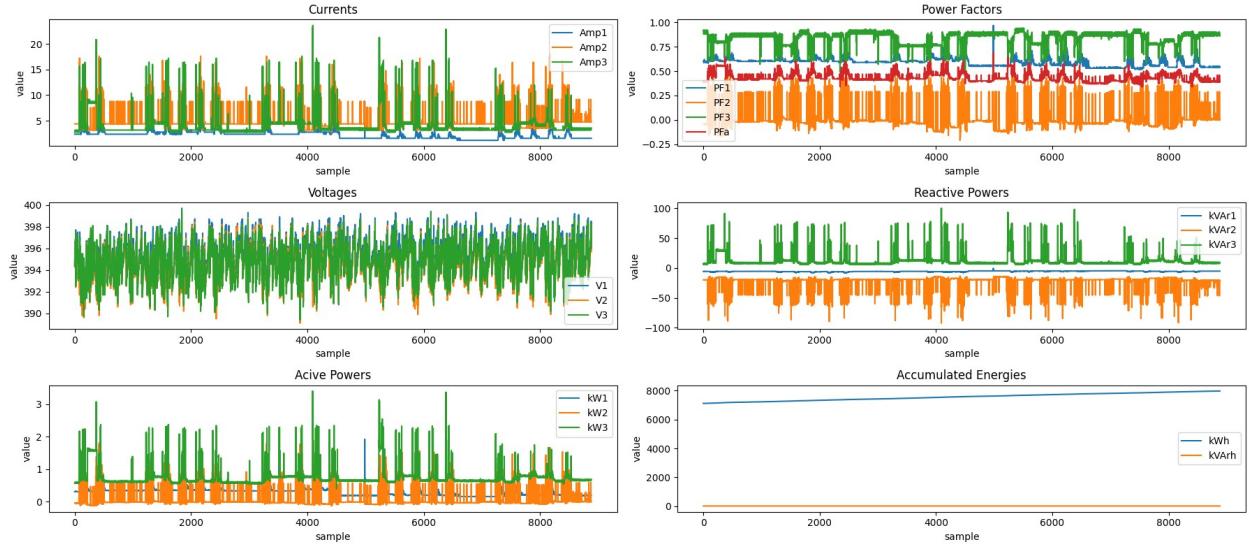
plt.subplot(3, 2, 3)
plt.plot(df[['V1', 'V2', 'V3']], label=['V1', 'V2', 'V3'])
plt.title('Voltages')
plt.xlabel("sample")
plt.ylabel('value')
plt.legend()

plt.subplot(3, 2, 4)
plt.plot(df[['kVAr1', 'kVAr2', 'kVAr3']], label=['kVAr1', 'kVAr2', 'kVAr3'])
plt.title('Reactive Powers')
plt.xlabel("sample")
plt.ylabel('value')
plt.legend()

plt.subplot(3, 2, 5)
plt.plot(df[['kW1', 'kW2', 'kW3']], label=['kW1', 'kW2', 'kW3'])
plt.title('Active Powers')
plt.xlabel("sample")
plt.ylabel('value')
plt.legend()

plt.subplot(3, 2, 6)
plt.plot(df[['kWh', 'kVArh']], label=['kWh', 'kVArh'])
plt.title('Accumulated Energies')
plt.xlabel("sample")
plt.ylabel('value')
plt.legend()

plt.tight_layout()
plt.show()
```



Notice how the spikes are the same or similar through all dependent quantities/columns of the same phase.

- **misalignments:**

I am not sure what does "misalignment" means exactly in this context. But my guess is that it means incorrect samples/rows ordering, missing timestamps or inconsistent time gaps, timestamps don't match datetimes, or unmatched relationships between values.

- The following code checks for **sample orders**:

```
# convert the datetime column's data type from string into
TimestampSeries in order to use .diff() directly.
diff_dt = pd.to_datetime(df['datetime'], format="%Y-%m-%d %I-%M-%S
%p")

print('number of unordered datetime values',
(diff_dt.diff().dt.total_seconds() < 0).sum())
print('number of unordered timestamp values', (df['timestamp'].diff()
< 0).sum())

number of unordered datetime values 0
number of unordered timestamp values 0
```

- **missing timestamps or inconsistent time gaps:** The following code shows that the time difference between each sample and the next one is for the vast majority between 05:01 and 05:02 or 30100 and 30200 ms:

```
print("df['timestamp'].diff()", df['timestamp'].diff()[1:11].values)
print("\ndiff_dt.diff():")
print(diff_dt.diff()[1:11])

df['timestamp'].diff(): [301512. 301433. 301490. 301528. 301372.
301673. 301485. 301505. 301417.]
```

```

301515.]

diff_dt.diff():
1    0 days 00:05:02
2    0 days 00:05:01
3    0 days 00:05:02
4    0 days 00:05:01
5    0 days 00:05:02
6    0 days 00:05:02
7    0 days 00:05:01
8    0 days 00:05:01
9    0 days 00:05:02
10   0 days 00:05:01
Name: datetime, dtype: timedelta64[ns]

```

This is perfect, 5 minutes is equal to 300000 ms! But now let's check whether or not there are any missing samples or inconsistent gaps:

```

dt_seconds = diff_dt.diff().dt.total_seconds().dropna() # the first
element is always nan

dt_lt_301 = dt_seconds[dt_seconds < 301]
print('datetime derivatives that have a value less than 301: ')
print('\nderivatives: ', dt_lt_301.values)
print('\nindices: ', list(dt_lt_301.index))
print('\ncount: ', len(dt_lt_301))
print('\nmin: ', dt_lt_301.min())
print('max: ', dt_lt_301.max())
print('mean: ', dt_lt_301.mean())

print('\n=====\\n')

dt_gt_302 = dt_seconds[dt_seconds > 302]
print('datetime derivatives that have a value greater than 302: ')
print('\nderivatives: ', dt_gt_302.values)
print('\nindices: ', list(dt_gt_302.index))
print('\ncount: ', len(dt_gt_302))
print('\nmin: ', dt_gt_302.min())
print('max: ', dt_gt_302.max())
print('mean: ', dt_gt_302.mean())

print('\\n=====\\n')

ts_ms = df['timestamp'].diff().dropna()

ts_lt_301000 = ts_ms[ts_ms < 301000]
print('timestamp derivatives that have a value less than 301000: ')
print('\nderivatives: ', ts_lt_301000.values)
print('\nindices: ', list(ts_lt_301000.index))
print('\ncount: ', len(ts_lt_301000))

```

```
print('\nmin: ', ts_lt_301000.min())
print('max: ', ts_lt_301000.max())
print('mean: ', ts_lt_301000.mean())

print('\n=====\\n')

ts_gt_302000 = ts_ms[ts_ms > 302000]
print('timestamp derivaties that have a value greater than 302000: ')
print('nderivaties: ', ts_gt_302000.values)
print('nindices: ', list(ts_gt_302000.index))
print('ncount: ', len(ts_gt_302000))
print('nmin: ', ts_gt_302000.min())
print('max: ', ts_gt_302000.max())
print('mean: ', ts_gt_302000.mean())

datetime derivaties that have a value less than 301:
derivaties: [300. 300. 300. 300. 300. 300. 300. 300.]

indices: [123, 1513, 1951, 3951, 3953, 4185, 5290, 5923]

count: 8

min: 300.0
max: 300.0
mean: 300.0

=====
datetime derivaties that have a value greater than 302:
derivaties: [316. 310. 308. 303. 313. 305. 307. 311. 310. 308. 311.
307. 303. 307.
311. 303. 312. 306. 303. 310. 308. 311. 307. 309. 309. 310. 309. 316.
310. 308. 314. 304. 315. 303. 315. 303. 303. 303. 309. 308. 304. 313.
310. 308. 306. 304. 314. 315. 303. 315. 303. 315. 303. 303. 316. 313.
305. 316. 317. 310. 308. 307. 311. 316. 303. 306. 312. 309. 309. 309.
309. 309. 309. 315. 303. 314. 304. 305. 312. 306. 303. 305. 313. 315.
303.]
```

indices: [176, 415, 416, 463, 642, 643, 875, 876, 1108, 1109, 1341, 1342, 1514, 1574, 1575, 1717, 1807, 1808, 1950, 2046, 2047, 2279, 2280, 2512, 2513, 2745, 2746, 2972, 3205, 3206, 3444, 3445, 3671, 3672, 3904, 3905, 3952, 3954, 4137, 4138, 4233, 4234, 4364, 4365, 4388, 4597, 4598, 4824, 4825, 5057, 5058, 5284, 5285, 5289, 5517, 5756, 5757, 5989, 6216, 6455, 6456, 6694, 6695, 6921, 6922, 7154, 7155, 7387, 7388, 7632, 7633, 7865, 7866, 8098, 8099, 8337, 8338, 8430, 8431, 8493, 8494, 8558, 8559, 8779, 8780]

count: 85

```
min: 303.0
max: 317.0
mean: 308.77647058823527
=====
timestamp derivaties that have a value less than 301000:
derivaties: [300372. 300779. 300582. 300975. 300883. 300694. 300951.
300864. 300903.
300776. 300957. 300869. 300746. 300952. 300805. 300958. 300961.
300788.
300964. 300782. 300193. 300995. 300683. 300557. 300946. 300823.
300997.
300984. 300981. 300928.]
```

indices: [123, 1398, 1451, 1479, 1500, 1513, 1518, 1555, 1557, 1627,
1671, 1951, 2557, 3111, 3887, 3946, 3951, 3953, 3981, 4002, 4185,
4259, 4435, 5290, 5600, 5923, 6309, 6454, 6972, 7686]

```
count: 30
=====
min: 300193.0
max: 300997.0
mean: 300821.6
=====
timestamp derivaties that have a value greater than 302000:
derivaties: [302012. 302574. 315669. 302494. 310285. 307615. 302064.
302074. 302068.
312806. 305343. 307035. 311188. 309976. 308045. 311925. 306269.
302170.
302008. 302010. 302189. 302141. 302308. 302104. 302115. 307463.
310567.
302499. 302166. 312377. 305855. 302006. 302074. 309577. 308428.
310178.
307831. 309237. 308727. 302049. 302337. 309560. 308530. 315869.
302140.
310168. 307785. 313876. 304191. 315195. 302867. 302012. 315514.
302541.
302148. 302107. 302160. 302051. 309207. 308679. 302321. 304089.
313443.
309888. 308253. 306447. 302083. 302350. 304297. 313608. 315907.
302379.
302033. 314774. 303373. 314999. 303160. 302291. 315820. 302038.
312749.
305359. 316284. 316728. 309930. 308060. 306318. 311659. 315294.
302987.
```

```

302028. 305459. 312453. 309563. 308630. 309121. 308878. 309144.
308959.
315779. 302376. 314399. 303845. 305798. 311772. 305876. 302576.
304978.
312897. 302127. 315084. 303354. 302015.]]

indices: [57, 122, 176, 177, 415, 416, 422, 463, 485, 642, 643, 875,
876, 1108, 1109, 1341, 1342, 1397, 1454, 1462, 1480, 1503, 1514, 1538,
1556, 1574, 1575, 1626, 1717, 1807, 1808, 1865, 1950, 2046, 2047,
2279, 2280, 2512, 2513, 2554, 2556, 2745, 2746, 2972, 2973, 3205,
3206, 3444, 3445, 3671, 3672, 3886, 3904, 3905, 3952, 3954, 3970,
4001, 4137, 4138, 4184, 4233, 4234, 4364, 4365, 4388, 4391, 4434,
4597, 4598, 4824, 4825, 4862, 5057, 5058, 5284, 5285, 5289, 5517,
5518, 5756, 5757, 5989, 6216, 6455, 6456, 6694, 6695, 6921, 6922,
6971, 7154, 7155, 7387, 7388, 7632, 7633, 7865, 7866, 8098, 8099,
8337, 8338, 8430, 8431, 8493, 8494, 8558, 8559, 8753, 8779, 8780,
8820]

count: 113

min: 302006.0
max: 316728.0
mean: 307101.6814159292

```

So far everything is perfect enough. All samples are ordered correctly and the difference in times are consistent. As it's shown above, some time differences are slightly bigger or shorter than normal. But this isn't an issue since these differences are very small and can just be neglected.

- **timestamps don't match datetimes:** The following code will check this one:

```

a = pd.to_datetime(df['datetime'], format="%Y-%m-%d %I-%M-%S%p").apply(lambda x: int(x.timestamp()))
a

0      1759276861
1      1759277163
2      1759277464
3      1759277766
4      1759278067
...
8877    1761953857
8878    1761954159
8879    1761954460
8880    1761954762
8881    1761955063
Name: datetime, Length: 8882, dtype: int64

b = df['timestamp'].apply(lambda x: int(x / 1000))
b

```

```
0      1759266061
1      1759266363
2      1759266664
3      1759266966
4      1759267267
       ..
8877    1761943057
8878    1761943359
8879    1761943660
8880    1761943962
8881    1761944263
Name: timestamp, Length: 8882, dtype: int64
```

At first glance, I thought that this huge differences between the two columns/features is a type of error or something. But after digging deeper, I noticed that all the differences are actually the same. There is exactly a 3 hours shift between all the values in these two columns. The following code illustrates this:

```
print('total differences:', ((a - b) != 0).sum(), '/', a.count())
print('unique differences:', (a - b).unique())

total differences: 8882 / 8882
unique differences: [10800]
```

This 3 hours shift is not a big deal actually. It is caused because the column "datetime" doesn't use a time zone, meanwhile, "timestamp" uses the Saudi time zone, i.e. UTC+3. This can be easily fixed by just adding or removing 3 hours from any column, or just use the same time zone for both.

Now we just showed that exactly all the values between these two columns are 100% consistent!

- **unmatched relationships between values:** Some physical quantities have relationships, for example, the power must increase if the current also increased but the voltage stayed constant. The following code will check these relationships:


```
[8882 rows x 26 columns]
```

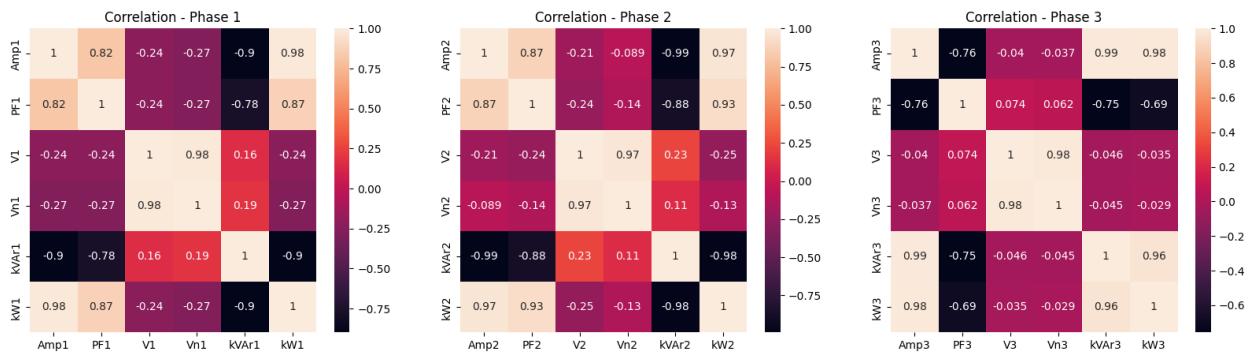
```
import seaborn as sns

plt.figure(figsize=(20, 5))

plt.subplot(1, 3, 1)
sns.heatmap(df[['Amp1', 'PF1', 'V1', 'Vn1', 'kVar1', 'kW1']].corr(),
            annot=True)
plt.title("Correlation - Phase 1")

plt.subplot(1, 3, 2)
sns.heatmap(df[['Amp2', 'PF2', 'V2', 'Vn2', 'kVar2', 'kW2']].corr(),
            annot=True)
plt.title("Correlation - Phase 2")

plt.subplot(1, 3, 3)
sns.heatmap(df[['Amp3', 'PF3', 'V3', 'Vn3', 'kVar3', 'kW3']].corr(),
            annot=True)
plt.title("Correlation - Phase 3")
plt.show()
```



Correlations Discussion:

- Amp vs. kW:** The graph shows strong positive correlation between these two features through all phases. This makes sense. If you looked at the graph that plots the voltages, you'll notice that they are constant but with noises. Thus, since the voltages are mostly constant, and only currents suffers from sudden spikes, the Active Power will be driven by the current heavily.
- Amp vs. PF:** This correlation is strong though all phases but the sign changes. After a fast research I did, I guess that this behaviour is possible in cases where the current affects the power efficiency heavily.

I didn't discuss the rest due to time shortage. But in a nutshell, these correlations are physically possible.

Q 1.4 Classify each telemetry key (continuous, categorical, boolean, etc.):

After inspecting the data file manually, I noticed that all columns/features have continuous values except for two, which are the timestamp, and datetime. The timestamp values are discrete integers, while datetime is a string.

Also, I think the following code illustrates this better:

```
df.dtypes
```

timestamp	int64
PF1	float64
kW1	float64
kVAr1	float64
kWh	float64
Amp1	float64
kVArh	float64
Hz	float64
Vn1	float64
Amp2	float64
V2	float64
Amp3	float64
V1	float64
Vn2	float64
V3	float64
Vn3	float64
kVAr3	float64
TkVAr	float64
kW2	float64
TkW	float64
PFa	float64
kVAr2	float64
PF3	float64
PF2	float64
kW3	float64
datetime	object
dtype:	object

Q 1.5 Briefly explain what the telemetry values represent operationally:

These values are collected from multiple sensors almost every 5 minutes. The sensors measures a lot of important physical quantities to grasp a strong understanding about the power consumption of a building. Some of these values suffer from noises such as the Voltages. But after inspecting a little bit, I guess this noise didn't affect the rest of quantities/features/columns that much. Also some values suffered from sudden high spikes such as currents. And I noticed how these same spikes got repeated in all the other quantities/values that depends on the current, such as powers for example. I guess such behaviour means that these anomalies aren't

caused by sensor errors, and I guess that these behaviours are caused can only be caused naturally or by something like humanitarian intervention because otherwise, the anomalies/spikes will only appear in one feature independently from the others. But this is not what happened. Unless if these dependent quantities/features/columns were not actually measured by sensors, but instead, they were calculated using algorithms after measuring the necessary quantities such as the voltage and current.

Another note is that the correlation between the physical quantities/features/columns are physical possible or correct. This means that the data still can be used for training models even tho it suffers from some noises and anomalies because when ever a quantity/feature faces a strange behavoir, all the other dependent quantities face similar behaviours with respect to their relationship. This means even if some values were incorrect, they still convey valuable information that correctly express the physical relationships between these quantities.

2 Data Pre-Processing

There are no missing or corrupted readings as I showed in the previous question. Thus, this will just be skipped.

```
# prepare a copy for preprocessing so we don't ruin the original data
processed_df = df.copy()
```

1. Encoding for non-numeric fields:

The only feature that isn't numeric is the "datetime". I prefer to just drop this column since I proved that it is consistent with timestamp. So we can just use timestamp instead. But here are two simple methods for encoding:

1.1 Using sklearn.preprocessing.LabelEncoder:

```
from sklearn.preprocessing import LabelEncoder

label_encoder = LabelEncoder()

processed_df['datetime'] =
label_encoder.fit_transform(processed_df['datetime'])
processed_df['datetime']

0      263
1      265
2      267
3      269
4      271
...
8877    8848
8878    8850
8879    8852
8880    8854
```

```
8881    8856
Name: datetime, Length: 8882, dtype: int64
```

1.2 Converting Datetime Into Seconds (implemented previously):

This is the method I'm using through the rest of code. If you want to use the first one, some steps later steps will not work as intended (e.g. time alignment) so you'll need to modify it yourself.

```
processed_df = df.copy()

processed_df['datetime'] = pd.to_datetime(processed_df['datetime'],
format="%Y-%m-%d %I-%M-%S %p").apply(lambda x: int(x.timestamp()))
processed_df['datetime']

0      1759276861
1      1759277163
2      1759277464
3      1759277766
4      1759278067
...
8877    1761953857
8878    1761954159
8879    1761954460
8880    1761954762
8881    1761955063
Name: datetime, Length: 8882, dtype: int64
```

2. Time alignment and resampling:

I showed previously how samples were collected regularly after approximately every 5 minutes. Thus, no much work is needed. But we can resample by scaling the timestamp column by a scalar if we want. One thing need to be done is to align the datetime column's values with the timestamp since they are 3 hours different and they also have different units now (the datetime got converted into seconds while timestamps is in ms):

```
# convert into seconds
processed_df['timestamp'] = processed_df['timestamp'].apply(lambda x:
int(x / 1000))

# remove the time zone affect
processed_df['datetime'] -= 10800

processed_df.head()

      timestamp  PF1    kW1   kVAr1      kWh  Amp1   kVArh     Hz    Vn1
Amp2 \
0  1759266061  0.6  0.308     -5.6  7123.6    2.4    8.33  59.94  227.8
4.4
```

```

1 1759266363 0.6 0.308 -5.6 7124.0 2.4 8.33 59.99 228.1
4.4
2 1759266664 0.6 0.312 -5.6 7124.0 2.4 8.33 60.00 228.4
4.4
3 1759266966 0.6 0.308 -5.6 7124.0 2.4 8.33 59.98 228.5
4.4
4 1759267267 0.6 0.308 -5.6 7124.0 2.4 8.33 60.00 228.4
4.4

... kVAr3 TkVAr kW2 TkW PFa kVAr2 PF3 PF2 kW3
datetime
0 ... 6.1 -19.3 -0.040 0.832 0.39 -19.8 0.92 -0.04 0.564
1759266061
1 ... 6.1 -19.3 -0.040 0.832 0.39 -19.8 0.92 -0.04 0.564
1759266363
2 ... 6.0 -19.4 -0.036 0.840 0.39 -19.8 0.92 -0.04 0.564
1759266664
3 ... 6.0 -19.5 -0.036 0.836 0.39 -19.9 0.92 -0.04 0.564
1759266966
4 ... 8.2 -17.1 -0.044 0.868 0.39 -19.7 0.88 -0.04 0.604
1759267267

[5 rows x 26 columns]

```

3. Pivoting data into a time-series structure:

```
processed_df = processed_df.set_index("datetime")
```

```
processed_df.head()
```

	timestamp	PF1	kW1	kVAr1	kWh	Amp1	kVArh	Hz		
Vn1 \ datetime	1759266061	1759266061	0.6	0.308	-5.6	7123.6	2.4	8.33	59.94	
227.8	1759266363	1759266363	0.6	0.308	-5.6	7124.0	2.4	8.33	59.99	
228.1	1759266664	1759266664	0.6	0.312	-5.6	7124.0	2.4	8.33	60.00	
228.4	1759266966	1759266966	0.6	0.308	-5.6	7124.0	2.4	8.33	59.98	
228.5	1759267267	1759267267	0.6	0.308	-5.6	7124.0	2.4	8.33	60.00	
228.4										
Amp2 \ datetime		Amp2	...	Vn3	kVAr3	TkVAr	kW2	TkW	PFa	kVAr2
1759266061	4.4	...	227.6	6.1	-19.3	-0.040	0.832	0.39	-19.8	

```

0.92
1759266363 4.4 ... 227.6 6.1 -19.3 -0.040 0.832 0.39 -19.8
0.92
1759266664 4.4 ... 227.9 6.0 -19.4 -0.036 0.840 0.39 -19.8
0.92
1759266966 4.4 ... 227.9 6.0 -19.5 -0.036 0.836 0.39 -19.9
0.92
1759267267 4.4 ... 228.2 8.2 -17.1 -0.044 0.868 0.39 -19.7
0.88

          PF2      kW3
datetime
1759266061 -0.04  0.564
1759266363 -0.04  0.564
1759266664 -0.04  0.564
1759266966 -0.04  0.564
1759267267 -0.04  0.604

[5 rows x 25 columns]

```

4. Outlier detection & Handling missing or corrupted readings:

I've already implemented an outlier detection algorithm in the previous question. I'll use the same method here, and I'll discuss three simple and easy to implement approaches for dealing with outliers. The first approach is to clip the extreme values, the second approach is to replace outliers with the mean value, and the third approach takes the values resulted from the second approach and replace the outlier in a window with the average.

I've tested the three approaches and got confused which one of the last two to pick. The last one looks better but I was afraid that it might remove some valuable information from the signal since I notice it smoothed it a little bit. But I tried it with a denoiser and the results were similar with the ones I got after using the second approach. Thus, I'll pick the last approach.

```

columns2clean = ['Amp1', 'Amp2', 'Amp3', 'Hz', 'PF1', 'PF2', 'PF3',
'PFA', 'TkVar', 'TkW', 'V1', 'V2', 'V3', 'Vn1', 'Vn2', 'Vn3', 'kVar1',
'kVar2', 'kVar3', 'kw1', 'kw2', 'kw3']

cliped = processed_df.copy()
averaged = processed_df.copy()

for column in columns2clean:
    Q1 = processed_df[column].quantile(0.25)
    Q3 = processed_df[column].quantile(0.75)
    IQR = Q3 - Q1

    # first approach
    cliped[column] = processed_df[column].clip(Q1 - 1.5 * IQR, Q3 +
1.5 * IQR)

```

```

# second approach
outliers = processed_df[(processed_df[column] < Q1 - 1.5 * IQR) | 
(processed_df[column] > Q3 + 1.5 * IQR)][column]
averaged.loc[outliers.index, column] = processed_df[column].mean()

# third approach
windowed = averaged.copy()

for column in columns2clean:
    roll = windowed[column].rolling(window=100, center=True,
min_periods=1)

    Q1 = roll.quantile(0.25)
    Q3 = roll.quantile(0.75)
    IQR = Q3 - Q1

    mask = (windowed[column] < Q1 - 1.5 * IQR) | (windowed[column] >
Q3 + 1.5 * IQR)

    rolling_mean = roll.mean().fillna(windowed[column].mean())
    windowed.loc[mask, column] = rolling_mean[mask]

plt.figure(figsize=(20, 8))

plt.subplot(3, 4, 1)
plt.plot(processed_df['Amp3'])
plt.xlabel("timestamp")
plt.title('Original Amp3')
plt.ylabel('value')

plt.subplot(3, 4, 2)
plt.plot(cliped['Amp3'])
plt.xlabel("timestamp")
plt.title('Cliped Amp3')
plt.ylabel('value')

plt.subplot(3, 4, 3)
plt.plot(averaged['Amp3'])
plt.xlabel("timestamp")
plt.title('Averaged Amp3')
plt.ylabel('value')

plt.subplot(3, 4, 4)
plt.plot(windowed['Amp3'])
plt.xlabel("timestamp")
plt.title('Windowed Amp3')
plt.ylabel('value')

plt.subplot(3, 4, 5)
plt.plot(processed_df['kVAr3'])

```

```
plt.xlabel("timestamp")
plt.title('Original kVar3')
plt.ylabel('value')

plt.subplot(3, 4, 6)
plt.plot(cliped['kVar3'])
plt.xlabel("timestamp")
plt.title('Cliped kVar3')
plt.ylabel('value')

plt.subplot(3, 4, 7)
plt.plot(averaged['kVar3'])
plt.xlabel("timestamp")
plt.title('Averaged kVar3')
plt.ylabel('value')

plt.subplot(3, 4, 8)
plt.plot(windowed['kVar3'])
plt.xlabel("timestamp")
plt.title('Windowed kVar3')
plt.ylabel('value')

plt.subplot(3, 4, 9)
plt.plot(processed_df['kW3'])
plt.xlabel("timestamp")
plt.title('Original kW3')
plt.ylabel('value')

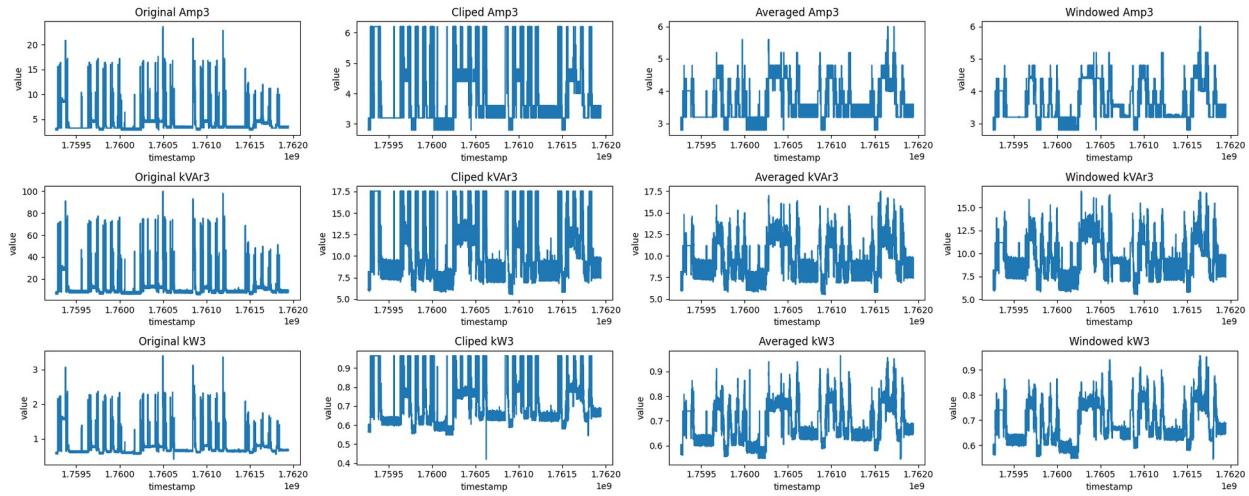
plt.subplot(3, 4, 10)
plt.plot(cliped['kW3'])
plt.xlabel("timestamp")
plt.title('Cliped kW3')
plt.ylabel('value')

plt.subplot(3, 4, 11)
plt.plot(averaged['kW3'])
plt.xlabel("timestamp")
plt.title('Averaged kW3')
plt.ylabel('value')

plt.subplot(3, 4, 12)
plt.plot(windowed['kW3'])
plt.xlabel("timestamp")
plt.title('Windowed kW3')
plt.ylabel('value')

plt.tight_layout()
plt.show()

processed_df = windowed
```



5. De-noising methods:

There are plenty of de-noising methods out there that I can use. But I'll just use the famous moving average filter.

I noticed after playing with the parameters for a while that applying the same filter but with smaller window size gives a lot smoother plots.

```
# moving average filter
old = processed_df.copy()

for i in range(10):
    processed_df[clean_columns] =
    processed_df[clean_columns].apply(lambda x: x.rolling(window=30,
    center=True).mean().ffill().bfill())

plt.figure(figsize=(16, 6))

plt.subplot(3, 2, 1)
plt.plot(old['V1'])
plt.xlabel("timestamp")
plt.title('V1 before')
plt.ylabel('value')

plt.subplot(3, 2, 2)
plt.plot(processed_df['V1'])
plt.xlabel("timestamp")
plt.title('V1 after')
plt.ylabel('value')

plt.subplot(3, 2, 3)
plt.plot(old['Amp3'])
plt.xlabel("timestamp")
plt.title('Amp3 before')
plt.ylabel('value')
```

```

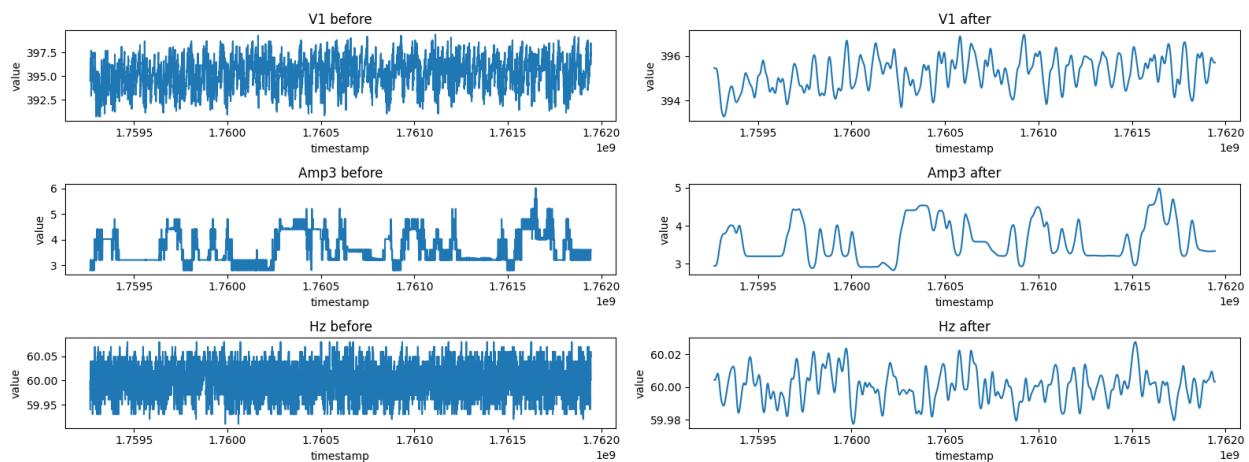
plt.subplot(3, 2, 4)
plt.plot(processed_df['Amp3'])
plt.xlabel("timestamp")
plt.title('Amp3 after')
plt.ylabel('value')

plt.subplot(3, 2, 5)
plt.plot(old['Hz'])
plt.xlabel("timestamp")
plt.title('Hz before')
plt.ylabel('value')

plt.subplot(3, 2, 6)
plt.plot(processed_df['Hz'])
plt.xlabel("timestamp")
plt.title('Hz after')
plt.ylabel('value')

plt.tight_layout()
plt.show()

```



6. Normalization / scaling:

As I showed previously, the scales of the different quantities are significantly various. These scale differences can cause several issues such as making it harder for models to train, producing unstable gradients, and it might even cause some issues such as saturated neurons at some model architectures.

There are several algorithms to apply scaling, but I'll just use a simple and fast approach. I'll normalize each column making its values all between 0.0 and 1.0 using the following code:

```

from sklearn.preprocessing import minmax_scale
processed_df = processed_df.apply(lambda x: minmax_scale(x,

```

```

feature_range=(0.0, 1.0)))

processed_df.describe()

      timestamp          PF1           kW1          kVAr1          kWh
\count  8882.000000  8882.000000  8882.000000  8882.000000  8882.000000
mean    0.499995    0.505413    0.451380    0.523114    0.522132
std     0.288724    0.254460    0.283134    0.266827    0.288213
min    0.000000    0.000000    0.000000    0.000000    0.000000
25%    0.249993    0.273379    0.163494    0.294773    0.277440
50%    0.499998    0.579734    0.548937    0.508790    0.539468
75%    0.749994    0.676263    0.641498    0.738986    0.779893
max    1.000000    1.000000    1.000000    1.000000    1.000000

      Amp1        kVArh          Hz          Vn1          Amp2 ...
\count  8882.000000  8882.0    8882.000000  8882.000000  8882.000000 ...
mean    0.422145    0.0    0.458155    0.510147    0.659508 ...
std     0.243750    0.0    0.188954    0.191520    0.185174 ...
min    0.000000    0.0    0.000000    0.000000    0.000000 ...
25%    0.171034    0.0    0.339414    0.371160    0.614523 ...
50%    0.488295    0.0    0.459500    0.508326    0.666531 ...
75%    0.619623    0.0    0.569257    0.655833    0.705581 ...
max    1.000000    0.0    1.000000    1.000000    1.000000 ...

      Vn3        kVAr3          TkVar          kW2          TkW
\count  8882.000000  8882.000000  8882.000000  8882.000000  8882.000000
mean    0.523878    0.402513    0.330442    0.492447    0.491748
std     0.204269    0.245151    0.205718    0.163924    0.248403
min    0.000000    0.000000    0.000000    0.000000    0.000000

```

25%	0.377313	0.207976	0.178806	0.387900	0.282732
50%	0.525059	0.341420	0.280172	0.520971	0.461696
75%	0.678485	0.630901	0.426867	0.590143	0.697036
max	1.000000	1.000000	1.000000	1.000000	1.000000
	PFa	kVAr2	PF3	PF2	kW3
count	8882.000000	8882.000000	8882.000000	8882.000000	8882.000000
mean	0.360255	0.401567	0.586244	0.674705	0.406687
std	0.184807	0.195645	0.256486	0.183197	0.236009
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.226322	0.283677	0.454309	0.572030	0.233533
50%	0.327871	0.434721	0.680411	0.714478	0.350594
75%	0.501854	0.521324	0.769334	0.793365	0.631270
max	1.000000	1.000000	1.000000	1.000000	1.000000
[8 rows x 25 columns]					

3 Time-Series Analysis

Q 3.1 Why did you choose this method?

I implemented an algorithm that relies on the Interquartile Range (IQR) Statistical method to detect outliers since it is stable, reliable, simple, easy to implement, I am experienced with it in college, and also I tried it and it worked out just great with this dataset. The algorithm also uses the moving average filter to smooth out the readings.

The way the algorithm works is by first detecting the outliers using IQR in all the values without using a window, it then replaces these values with an average. Then it repeats the first step but with using a window instead of applying the replacement mechanism to all values at ones. Finally, it will apply the moving average denoising filter to smooth the signal and remove further spikes or reduce them.

Q 3.2 How would it handle sudden spikes, sensor errors, or abnormal behavior?

- **sudden spikes:** It handles this by replacing the outliers with the average value twice (once without a window and then with a window), and then it uses a

denoising filter to smooth any remaining minor fluctuations after replacement. After these steps, we end up with removed spikes or reduced, resulting in a way more cleaner and stable signal.

- **sensor errors & abnormal behavior:** Noises, extreme values, or physically impossible values, all of these issues will be easily dealt with as follows: The noise will be removed using the denoising filter, the extreme values will get removed or reduced by replacing them with the average value, and the physically impossible values will simply be detected as outliers and replaced.

Q 3.3 What features or windows would you use?

- **features:**

If by "features" you mean what columns would I preprocess, then I picked these 22 features: Amp1, Amp2, Amp3, Hz, PF1, PF2, PF3, PFa, TkVAr, TkW, V1, V2, V3, Vn1, Vn2, Vn3, kVAr1, kVAr2, kVAr3, kW1, kW2, kW3

These are the features which are suffering from abnormal behaviours. Of course, the datetime and timestamp don't count. And the cumulative features, kVArh and kWh, are very stable and smooth and don't need to be further cleaned.

- **windows:**

For detecting outliers I didn't use a window at first, I applied the calculations at all the values. After that I repeated the same process but with using a window of size 100. But for the denoising filter, I used a window of size 30 and applied the filter 10 times. I reached this approach after several tests.

The following is the complete code implementation for this approach:

```
def data_cleaner(data):  
    df = data.copy()  
  
    columns2clean = ['Amp1', 'Amp2', 'Amp3', 'Hz', 'PF1', 'PF2',  
    'PF3', 'PFa', 'TkVAr', 'TkW', 'V1', 'V2', 'V3', 'Vn1', 'Vn2', 'Vn3',  
    'kVAr1', 'kVAr2', 'kVAr3', 'kW1', 'kW2', 'kW3']  
  
    # outliers replacement  
    for column in columns2clean:  
        Q1 = df[column].quantile(0.25)  
        Q3 = df[column].quantile(0.75)  
        IQR = Q3 - Q1  
  
        outliers = df[(df[column] < Q1 - 1.5 * IQR) | (df[column] > Q3  
+ 1.5 * IQR)][column]  
        df.loc[outliers.index, column] = df[column].mean()  
  
        roll = df[column].rolling(window=100, center=True,  
        min_periods=1)
```

```

Q1 = roll.quantile(0.25)
Q3 = roll.quantile(0.75)
IQR = Q3 - Q1

mask = (df[column] < Q1 - 1.5 * IQR) | (df[column] > Q3 + 1.5
* IQR)

rolling_mean = roll.mean().fillna(df[column].mean())
df.loc[mask, column] = rolling_mean[mask]

# de-noising
for i in range(10):
    df[columns2clean] = df[columns2clean].apply(lambda x:
x.rolling(window=30, center=True).mean().ffill().bfill())

return df

```

4 IoT System Logic & Operational Thinking

At a certain timestamp, the system reports the following values:

kW1	0.0
Amp1	0.2
PF1	0.60
Vn1	228.0
Hz	60.00
kWh	continues increasing normally

Using only the telemetry keys available in the attached file:

1. Is this behavior normal or abnormal?

Hz and Vn1 are totally normal, actually, they are equal to their average values. kWh is normally increasing even though kW1 is 0. This might happen due to the fact that this is a three phases building. This means that kW2 and kW3 might not be zero. So far. Now when it comes to Amp1 and kW1, although it is normal for Amp1 and kW1 to decrease together since they are directly proportioned, but this behavior didn't happen before in the dataset. For Amp1, the lowest value it had in the dataset is 1.2, and for kW1, is 0.16. I didn't understand the behavior of PF1, it has the average value which is 0.6. But PF1 is directly proportioned with both kW1 and Amp1 as the dataset statistics that I calculated before showed. This might happen if kVar1 decreased too, but I have no clue about this.

My final answer is while this behavior might be physically possible, it is still abnormal.

2. Provide an operational interpretation of what might cause this reading.

Since the voltage value is normal and the frequency too, this means that the phase is connected to power, but no active power is being drawn by devices.

3. Discuss at least 3 possible explanations, such as measurement delays, current leakage, low-load conditions, sensor inaccuracies, or accumulated energy behavior.

1. **Current Leakage in Idle Phase:** This behavior might occur if there are no devices drawing active power, but there is still some tiny current leakage.
2. **Capacitive Load:** If there are no devices drawing active power, and there is a capacitive load, then we'll see such a behavior. The active power is 0 because no device is consuming it, and the small current is produced by a capacitive load. The value of the reactive power (kVar1) is always negative in the dataset, which indicates an existence of a capacitive load. This is why I think this explanation is the most probable.
3. **Sensor Inaccuracies:** The abnormal values of Amp1 and kW1 might be errors caused by the sensors. I think this explanation is the most unlikely. Because through all the samples that being collected each 5 minutes for 31 days, there is no a single instance of any behavior similar to this. Also the only abnormal values happen to be physically related, and they correctly matched their physical relationship. These notes increase the chance of these values being real and correctly detected by sensors.

5 Predictive Modeling

- **Pseudocode:**

Here is a pseudocode for an LSTM model that takes the water consumption for 60 minutes (this is an arbitrary value) as an input, then predicts for the next 15 minutes:

```
#Loading and cleaning data
data = download("water consumption dataset")

data = handle_missing_values(data)
data = sort_by_timestamp(data)
data = handle_outliers(data)
data = denoise(data)
data = normalize(data)

#Feature engineering
data["day_of_week"] = extract_days(data)

#Defining the prediction window
INPUT_WINDOW = 60 # minutes
PREDICTION_WINDOW = 15 # minutes
```

```

func create_sequences(data)
    sequences = []
    targets = []

    for i in range(INPUT_WINDOW, len(data) - FORECAST_HORIZON)
        sequences.append(data[i - INPUT_WINDOW : i])
        labels.append(data[i + FORECAST_HORIZON])

    return sequences, labels

x, y = create_sequences(data)

#Splitting data (train/test)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size =
0.3) # split 30% of the data for testing, and 70% for training

#Model selection (e.g., LSTM, Random Forest, etc.)
model = LSTM(inp_size, out_size, num_layers=layers) # assuming
inp_size, out_size, and layers are defined

model.fit(loss='MSE', optimizer='AdamW', lr=1e-1, X_train, y_train,
epochs=100, batch_size=128)

#Prediction logic
print(model.predict(X_test)) #display predictions

```

- **Full Executable Code:**

This code implementation is slightly different. It's basically a model implementation that takes the water consumption measurements for the last 7 days as input and then predict the consumption for the next day.

1. The Dataset (Fiviple water consumption dataset):

I spent some time searching and exploring free datasets that I can use. I found few good datasets but unfortunately they had restricted access permissions. And I also found open-sourced datasets, but they all were very bad and almost useless in my opinion. After almost giving up on this, I suddenly got the idea of making my own water consumption dataset! And I made it very similar to any raw datasets so it still need preprocessing and feature engineering or else the model will perform bad. I didn't cheat, I didn't make easy for me.

It is a tabular dataset about the water consumption in one year for 5 people that live together in one apartment. Each row contains 5 measurements collected in one day. These measurements are: the lowest temperature, highest temperature, lowest humidity, highest humidity, and total water consumption. There are also some hidden information can be extracted from these measurements. For example, all the 5 measurements change following some trends, figuring out these trend can significantly improve the model performance and ease training process. Also, the fact that there are trends and the data isn't totally random raise the need of some model architecture that has some type of memory (e.g. LSTM, and GRU). But this need can be neglected

by simply giving an MLP (for example) multiple rows at once by flattening them together as if they were one row. In this way, the model will still be able to perform good enough at this dataset.

The dataset is public and open-source, and it can be found here:

<https://www.kaggle.com/datasets/firasghazialsadiq/fiviple-water-consumption-dataset>

1.1 Downloading The Dataset:

```
import kagglehub
import os

path = kagglehub.dataset_download("firasghazialsadiq/fiviple-water-
consumption-dataset")

path = os.path.join(path, 'water consumption dataset')
path

c:\Users\NEMFi\AppData\Local\Programs\Python\Python311\Lib\site-
packages\tqdm\auto.py:21: TqdmWarning: IPython not found. Please
update jupyter and ipywidgets. See
https://ipywidgets.readthedocs.io/en/stable/user_install.html
    from .autonotebook import tqdm as notebook_tqdm

Warning: Looks like you're using an outdated `kagglehub` version
(installed: 0.3.12), please consider upgrading to the latest version
(0.3.13).

'C:\\\\Users\\\\NEMFi\\\\.cache\\\\kagglehub\\\\datasets\\\\firasghazialsadiq\\\\
fiviple-water-consumption-dataset\\\\versions\\\\5\\\\water consumption
dataset'
```

1.2 Reading & Exploring The Dataset:

You'll find the following files inside the dataset's folder:

```
fiviple-water-consumption-dataset/
├── test.csv
├── train.csv
└── val.csv

os.listdir(path)
['test.csv', 'train.csv', 'val.csv']

import pandas

# read the datasets
train_df = pandas.read_csv(os.path.join(path, 'test.csv'))
test_df = pandas.read_csv(os.path.join(path, 'train.csv'))
val_df = pandas.read_csv(os.path.join(path, 'val.csv'))
```

```

# compare them
print('total number of rows in the training dataset:', len(train_df))
print('training dataset columns:', list(train_df.columns))

print('\ntotal number of rows in the testing dataset:', len(test_df))
print('testing dataset columns:', list(test_df.columns))

print('\ntotal number of rows in the validation dataset:',
len(val_df))
print('validation dataset columns:', list(val_df.columns))

total number of rows in the training dataset: 365
training dataset columns: ['low temp', 'high temp', 'low humidity',
'high humidity', 'water consumption']

total number of rows in the testing dataset: 365
testing dataset columns: ['low temp', 'high temp', 'low humidity',
'high humidity', 'water consumption']

total number of rows in the validation dataset: 365
validation dataset columns: ['low temp', 'high temp', 'low humidity',
'high humidity', 'water consumption']

plt.figure(figsize=(20, 2))

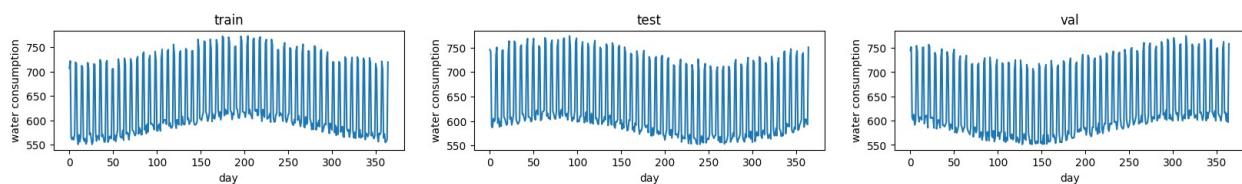
plt.subplot(1, 3, 1)
plt.plot(train_df['water consumption'])
plt.xlabel("day")
plt.title('train')
plt.ylabel('water consumption')

plt.subplot(1, 3, 2)
plt.plot(test_df['water consumption'])
plt.xlabel("day")
plt.title('test')
plt.ylabel('water consumption')

plt.subplot(1, 3, 3)
plt.plot(val_df['water consumption'])
plt.xlabel("day")
plt.title('val')
plt.ylabel('water consumption')

plt.show()

```



```
train_df.head()
```

	low	temp	high	temp	low	humidity	high	humidity	water consumption
0	-4.8		5.4		30.5		52.5		707.1
1	-4.7		5.4		30.7		53.0		721.9
2	-4.9		5.3		31.7		52.9		565.5
3	-4.8		5.5		30.3		52.7		560.7
4	-5.0		5.3		30.4		51.0		565.5

1.3 Custom Dataset Class & Preprocessing:

I created a custom dataset class to simplify the code later on and make dealing with data more flexible and smoother. Also I scaled the values to be in the range [-1, 1]. I came up with this after analyzing the performance and applying several tests. I managed to reach satisfying results without using other preprocessing techniques. Thus, I'll stop here.

```
from torch.utils.data import Dataset
from sklearn.preprocessing import minmax_scale
import torchvision.transforms as transforms

class WaterConsumptionDataset(Dataset):
    def __init__(self, input_window, path, split):
        self.input_window = max(1, input_window)
        df = pandas.read_csv(os.path.join(path, split + '.csv'))
        self.x = df
        # scaling
        self.x = self.x.apply(lambda x: minmax_scale(x,
        feature_range=(-1.0, 1.0)))
        self.y = df['water consumption']
        # to convert from numpy.array into torch.Tensor
        self.converter = transforms.ToTensor()

    def __len__(self):
        return len(self.x) - self.input_window

    def __getitem__(self, idx):
        if idx >= len(self):
            return
        return self.converter(self.x.iloc[idx : idx +
        self.input_window].values).squeeze(0), self.y[idx + self.input_window]

train_dataset = WaterConsumptionDataset(7, path, 'train')
test_dataset = WaterConsumptionDataset(7, path, 'train')
```

```
x, y = train_dataset[1]

print(x.shape)
print(y)

torch.Size([7, 5])
750.5
```

1.4 DataLoader:

To train the model faster, I used `torch.utils.data.DataLoader` which enables me to retrieve data in batches instead of processing them slowly one by one. I also shuffled the training dataset to push the model into understanding the data better.

```
from torch.utils.data import DataLoader

train_dataloader = DataLoader(train_dataset, batch_size=128,
shuffle=True)
test_dataloader = DataLoader(test_dataset, batch_size=1,
shuffle=False)

X, Y = next(iter(train_dataloader))

print(X.shape)
print(Y.shape)

torch.Size([128, 7, 5])
torch.Size([128])
```

2. The Model:

I created a model that is composed of 1 dense layer, 1 LSTM layer, and 2 Dense layers, respectfully. I got to this structure after analizing the performance and conducting several tests with different structures. I also tried to use activation functions such as `relu`, but the results were worse. And I also tried to apply dropout and it didn't help. Finally, I ended up with this structure which actually performed great.

```
import torch.nn as nn

class LSTMPredictor(nn.Module):
    def __init__(self, inp_size=5):
        super(LSTMPredictor, self).__init__()

        self.fc1 = nn.Linear(inp_size, 15)
        self.lstm = nn.LSTM(15, 30, num_layers=1, batch_first=True)
        self.fc2 = nn.Linear(30, 15)
        self.fc3 = nn.Linear(15, 1)

    def forward(self, X):
```

```

        X = self.fc1(X)
        X, _ = self.lstm(X)
        X = self.fc2(X[:, -1])
        X = self.fc3(X)

    return X

import torch

model = LSTMPredictor()

# run the model on the gpu if there is one
# this speeds up the training a lot
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
model.to(device)

LSTMPredictor(
    (fc1): Linear(in_features=5, out_features=15, bias=True)
    (lstm): LSTM(15, 30, batch_first=True)
    (fc2): Linear(in_features=30, out_features=15, bias=True)
    (fc3): Linear(in_features=15, out_features=1, bias=True)
)

```

3. Choosing The Optimizer and Criterion:

For the criterion/loss function, I used the mean squared error (MSE). Since this is a regression problem, MSE is the default choice here. But when it comes to the optimization method, I arbitrary chose AdamW. I came up with this learning rate value after testing several other values.

```

import torch.optim as optim

criterion = nn.MSELoss()
optimizer = optim.AdamW(model.parameters(), lr=1e-1)

```

4. Training:

I used the root mean squared error (RMSE) metric to display the training progress using a more interpretable values unlike MSE.

```

import matplotlib.pyplot as plt

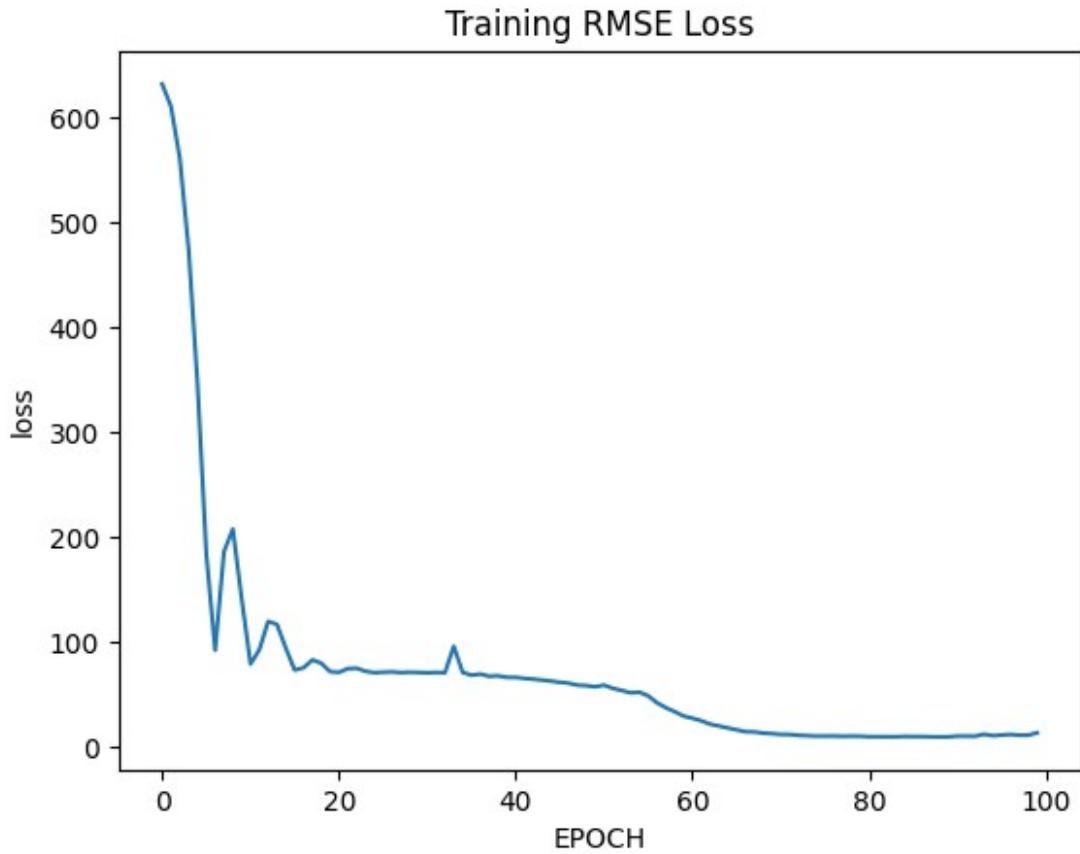
def train_model():
    train_losses = []

    model.train()

    for epoch in range(100):
        avg_loss = 0.0

```

```
for X, targets in train_dataloader:  
    X, targets = X.to(device).to(torch.float32),  
    targets.to(device).to(torch.float32)  
  
    optimizer.zero_grad()  
  
    predictions = model(X)  
  
    loss = criterion(predictions.reshape(-1), targets)  
    avg_loss += loss.sqrt().item()  
  
    loss.backward()  
    optimizer.step()  
  
    avg_loss /= len(train_dataloader)  
    train_losses.append(avg_loss)  
  
plt.plot(train_losses)  
plt.xlabel("EPOCH")  
plt.ylabel("loss")  
plt.title("Training RMSE Loss")  
plt.show()  
  
train_model()
```



5. Testing:

```
def test_model():
    from sklearn.metrics import root_mean_squared_error

    avg_loss = 0.0

    predictions = []
    targets = []

    model.eval()
    for X, target in test_dataloader:
        X, target = X.to(device).to(torch.float32),
        target.to(device).to(torch.float32)

        optimizer.zero_grad()

        prediction = model(X)

        predictions.append(prediction.item())
        targets.append(target.item())

        loss = root_mean_squared_error(target.cpu().detach().numpy(),
            prediction.reshape(-1).cpu().detach().numpy())
```

```

    avg_loss += loss
avg_loss /= len(test_dataloader)

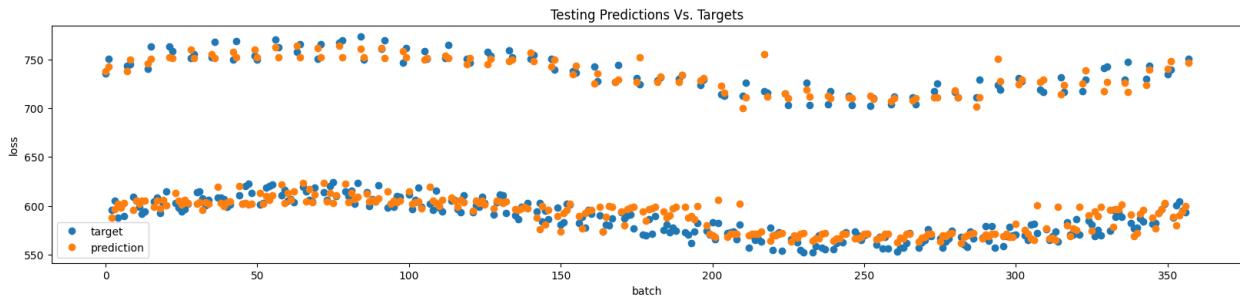
print('Average Test RMSE Loss:', avg_loss)

plt.figure(figsize=(20, 4))
plt.plot(targets, 'o', label='target', )
plt.plot(predictions, 'o', label='prediction')
plt.xlabel("batch")
plt.ylabel("loss")
plt.title("Testing Predictions Vs. Targets")
plt.legend()
plt.show()
plt.show()

test_model()

```

Average Test RMSE Loss: 8.441184443468488



6 Creativity & Feature Engineering

Although the model performed well, the performance can still be improved with some feature engineering techniques. At first, I tried to add an additional column to the inputs that represent the water consumption trend (`df['water_consumption'].diff()`) but it didn't help. After that I added two features, the first indicates whether a specific day is a weekend or not, the second one represents the number of the day (i.e. 1,2,3...365). With only these two additional features, it became much easier for the model to understand and predict the trends. The dataset contains minor noises, but with these two features, the model managed to ignore the noises and produce smooth and precise curves as shown below.

I only needed two lines of code to add these features. Thus, the dataset class is almost identical to the previous one.

```

class WaterConsumptionDataset(Dataset):
    def __init__(self, input_window, path, split):
        self.input_window = max(1, input_window)
        df = pandas.read_csv(os.path.join(path, split + '.csv'))

```

```

    self.x = df

    # feature engineering
    self.x['weekend'] = [1 if (d % 7) == 1 or (d % 7) == 2 else 0
for d in range(1, 366)] # 1
    self.x['day'] = [d for d in range(1, 366)] # 2

    self.x = self.x.apply(lambda x: minmax_scale(x,
feature_range=(-1.0, 1.0)))

    self.y = df['water consumption']

    self.converter = transforms.ToTensor()

def __len__(self):
    return len(self.x) - self.input_window

def __getitem__(self, idx):
    if idx >= len(self):
        return

    return self.converter(self.x.iloc[idx : idx +
self.input_window].values).squeeze(0), self.y[idx + self.input_window]

# run these again:

train_dataset = WaterConsumptionDataset(7, path, 'train')
test_dataset = WaterConsumptionDataset(7, path, 'train')

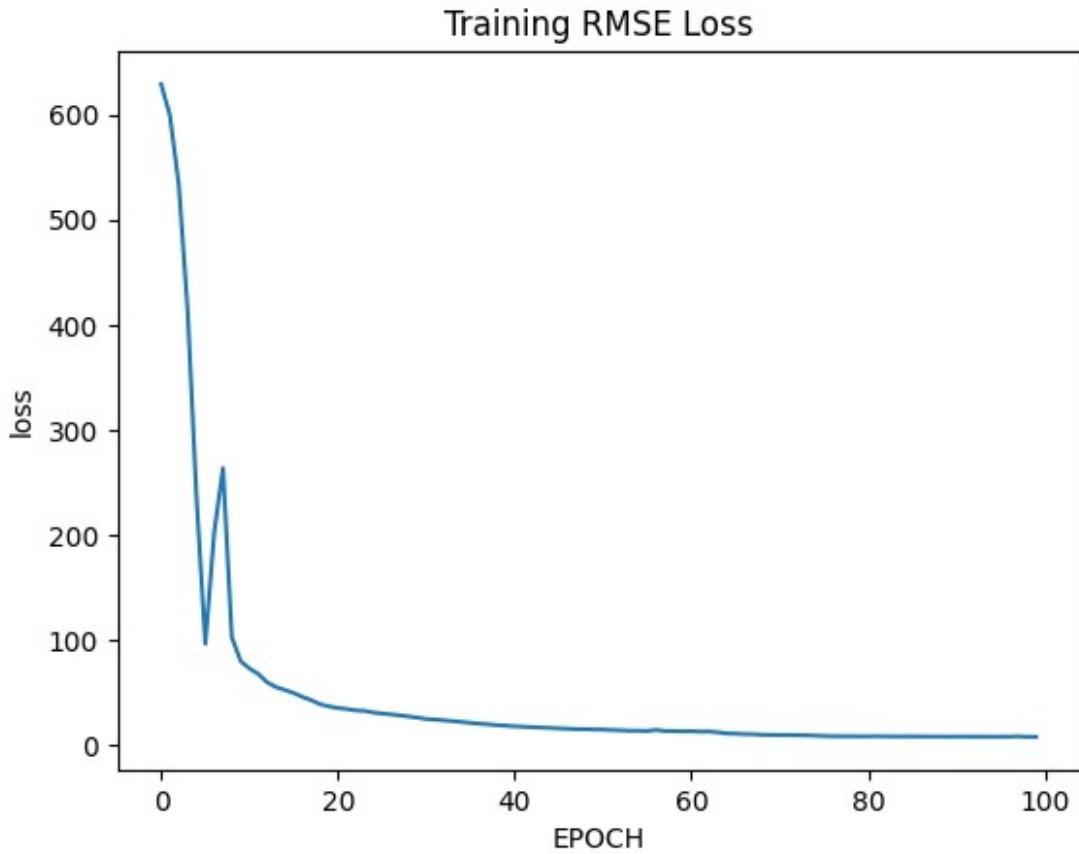
train_dataloader = DataLoader(train_dataset, batch_size=128,
shuffle=True)
test_dataloader = DataLoader(test_dataset, batch_size=1,
shuffle=False)

model = LSTMPredictor(7)
model.to(device)

optimizer = optim.AdamW(model.parameters(), lr=1e-1)

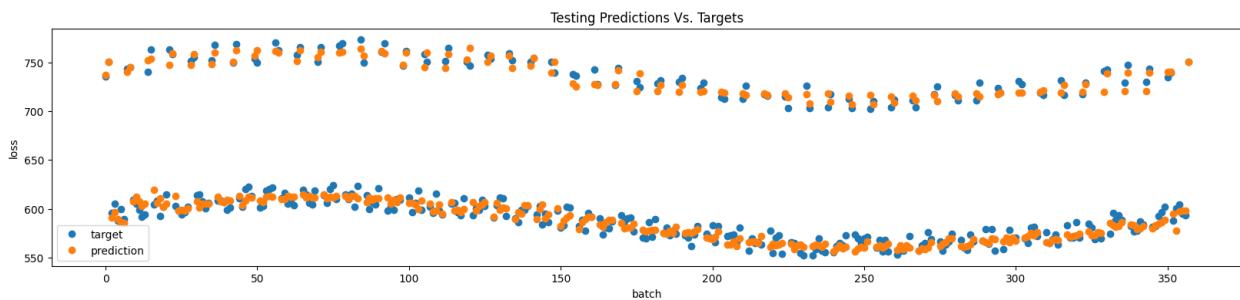
train_model()

```



```
test_model()
```

```
Average Test RMSE Loss: 6.236400582936889
```



Critical Issue:

I faced an annoying issue while training the model. Sometimes, it reached good results as the one seen here, but in other times it fails. This happens because Pytorch initializes the model's weights using some techniques, but these techniques are affected by some randomness. Usually, this randomness isn't an issue, but unfortunately, it is here. This issue became rarer to happen after using the engineered features, but it still exists. So, you might just need to reinitialize the model again if this code didn't show similar results. An easy fix to this is to just use a constant seed value that produces good models. But I got tired and I'll just stop here.