In [318…

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import warnings
warnings.filterwarnings('ignore')
```

In [319…

```
df=pd.read_csv('https://raw.githubusercontent.com/wri/global-power-plant-database/master/source_databases_csv/dat
```

In [320…

```
df.head(10)
```

Out[320…

	country	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	other_fuel2	...	year_of_capacity
0	IND	India	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	NaN	NaN	...	
1	IND	India	ADITYA CEMENT WORKS	WRI1019881	98.0	24.7663	74.6090	Coal	NaN	NaN	...	
2	IND	India	AES Saurashtra Windfarms	WRI1026669	39.2	21.9038	69.3732	Wind	NaN	NaN	...	
3	IND	India	AGARTALA GT	IND0000001	135.0	23.8712	91.3602	Gas	NaN	NaN	...	
4	IND	India	AKALTARA TPP	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	NaN	...	
5	IND	India	AKRIMOTA LIG	IND0000003	250.0	23.7689	68.6447	Coal	Oil	NaN	...	
6	IND	India	ALIYAR	IND0000004	60.0	10.4547	77.0078	Hydro	NaN	NaN	...	
7	IND	India	ALLAIN DUHANGAN	IND0000005	192.0	32.2258	77.2070	Hydro	NaN	NaN	...	
8	IND	India	ALMATTI DAM	IND0000006	290.0	16.3300	75.8863	Hydro	NaN	NaN	...	
9	IND	India	AMAR KANTAK	IND0000007	210.0	23.1642	81.6373	Coal	Oil	NaN	...	

10 rows × 27 columns

In [321…

```
df.shape
```

Out[321… (907, 27)

In [322…

```
df.dtypes
```

Out[322…

```
country                  object
country_long             object
name                     object
gppd_idnr                object
capacity_mw             float64
latitude                float64
longitude               float64
primary_fuel             object
other_fuel1              object
other_fuel2              object
other_fuel3             float64
commissioning_year      float64
owner                    object
source                   object
url                      object
geolocation_source       object
wepp_id                 float64
year_of_capacity_data   float64
generation_gwh_2013     float64
generation_gwh_2014     float64
generation_gwh_2015     float64
generation_gwh_2016     float64
generation_gwh_2017     float64
generation_gwh_2018     float64
generation_gwh_2019     float64
generation_data_source   object
```

```
estimated_generation_gwh    float64
dtype: object
```

In [323…
```
df.isnull().sum()
```

Out[323…
```
country                       0
country_long                  0
name                          0
gppd_idnr                     0
capacity_mw                   0
latitude                     46
longitude                    46
primary_fuel                  0
other_fuel1                 709
other_fuel2                 906
other_fuel3                 907
commissioning_year          380
owner                       565
source                        0
url                           0
geolocation_source           19
wepp_id                     907
year_of_capacity_data       388
generation_gwh_2013         907
generation_gwh_2014         509
generation_gwh_2015         485
generation_gwh_2016         473
generation_gwh_2017         467
generation_gwh_2018         459
generation_gwh_2019         907
generation_data_source      458
estimated_generation_gwh    907
dtype: int64
```

In [324…
```
##
```

In [325…
```
##
```

In [326…
```
# Checking number of unique values in each columns
count = 1
for x in df:
    print(f'{count}. {x}: {df[x].nunique()}')
    print(f'{df[x].value_counts()}', end = '\n----------\n\n' )
    count += 1
```

```
1. country: 1
IND    907
Name: country, dtype: int64
----------

2. country_long: 1
India    907
Name: country_long, dtype: int64
----------

3. name: 907
Sadeipali - REHPL Solar Power Plant    1
SANGLI MIRAJ BIOMASS                   1
BARADARHA TPP                          1
MAHATMA SUGAR                          1
SOUTHERN REPL.                         1
                                      ..
CHANDA CEMENT WORKS                    1
MAIHAR CEMENT PLANT                    1
MOHAMAD PUR                            1
Kathauti 2 Solar Power Plant           1
TASHIDING                              1
Name: name, Length: 907, dtype: int64
----------

4. gppd_idnr: 907
IND0000381    1
IND0000331    1
IND0000186    1
WRI1026092    1
```

```
WRI1026761    1
              ..
WRI1026115    1
IND0000417    1
IND0000524    1
IND0000256    1
IND0000471    1
Name: gppd_idnr, Length: 907, dtype: int64
----------

5. capacity_mw: 361
5.0       39
10.0      22
15.0      20
600.0     20
1200.0    19
          ..
192.0      1
27.3       1
26.4       1
68.8       1
19.7       1
Name: capacity_mw, Length: 361, dtype: int64
----------

6. latitude: 836
24.1917    3
19.0004    3
15.2615    2
13.2450    2
11.5336    2
           ..
16.4994    1
9.0870     1
20.9099    1
17.2387    1
16.5973    1
Name: latitude, Length: 836, dtype: int64
----------

7. longitude: 827
71.6917    4
72.8983    3
81.2875    3
75.8988    3
71.6918    3
           ..
77.9576    1
91.8114    1
80.1264    1
76.1137    1
79.5748    1
Name: longitude, Length: 827, dtype: int64
----------

8. primary_fuel: 8
Coal       258
Hydro      251
Solar      127
Wind       123
Gas         69
Biomass     50
Oil         20
Nuclear      9
Name: primary_fuel, dtype: int64
----------

9. other_fuel1: 3
Oil             195
Gas               2
Cogeneration      1
Name: other_fuel1, dtype: int64
----------

10. other_fuel2: 1
Oil    1
Name: other_fuel2, dtype: int64
----------

11. other_fuel3: 0
Series([], Name: other_fuel3, dtype: int64)
----------
```

```
12. commissioning_year: 73
2015.0    28
2013.0    25
2012.0    23
2016.0    19
2010.0    18
          ..
1958.0     1
1949.0     1
1954.0     1
1956.0     1
1927.0     1
Name: commissioning_year, Length: 73, dtype: int64
----------

13. owner: 280
Jk Cement ltd                                   4
Acc Acc ltd                                     4
Sterling Agro Industries ltd.                   4
Powerica Limited                                3
Shri Ssk ltd                                    3
                                               ..
Godavari Mills ltd                              1
Frost International Limited                     1
Solairedirect Projects India Private Limited    1
Saidham Overseas Private Limited                1
West Coast Paper Mills Ltd.                     1
Name: owner, Length: 280, dtype: int64
----------

14. source: 191
Central Electricity Authority               519
CDM                                         124
Lancosola                                    10
National Renewable Energy Laboratory          8
National Thermal Power Corporation (NTPC)     6
                                           ...
Real Estate e                                 1
EMC Limited                                   1
Lokmangal Lokmangal group                     1
Maral Overseas ltd                            1
West Coast Paper Mills Ltd.                   1
Name: source, Length: 191, dtype: int64
----------

15. url: 304
http://www.cea.nic.in/
519
http://www.lancosolar.com/pdfs/rajasthan-pv-project-details.pdf
7
http://www.ntpc.co.in
6
http://viainfotech.biz/Biomass/theme5/document/green_market/REC-project-list.pdf
5
http://energy.rajasthan.gov.in/content/dam/raj/energy/common/Details%20of%20commissioned%20Solar%20Projects%20.pd
f                              4

...
https://cdm.unfccc.int/filestorage/w/m/64TXH0Y1V9ZCISBKO3F758PEQNUJDR.pdf/PDD__V-2_5_19_10_2012.pdf?t=akh8b2pkZTF
tfDAP0pu4sjZSao0P-GV-Qzqn      1
http://documents.worldbank.org/curated/en/442061468041961880/pdf/multi-page.pdf
1
http://www.tradeindia.com/Seller-6835496-Shri-Dudhganga-Vedganga-SSK-Ltd-/
1
http://www.thoratsugar.com/
1
https://cdm.unfccc.int/Projects/DB/LRQA%20Ltd1346322352.66/view
1
Name: url, Length: 304, dtype: int64
----------

16. geolocation_source: 3
WRI                                   765
Industry About                        119
National Renewable Energy Laboratory    4
Name: geolocation_source, dtype: int64
----------

17. wepp_id: 0
Series([], Name: wepp_id, dtype: int64)
----------

18. year_of_capacity_data: 1
```

```
2019.0    519
Name: year_of_capacity_data, dtype: int64
----------

19. generation_gwh_2013: 0
Series([], Name: generation_gwh_2013, dtype: int64)
----------

20. generation_gwh_2014: 371
0.00000       28
6803.31250     1
4735.13000     1
145.81400      1
2022.57000     1
              ..
6224.00000     1
268.48085      1
1255.73200     1
164.32425      1
1153.65300     1
Name: generation_gwh_2014, Length: 371, dtype: int64
----------

21. generation_gwh_2015: 396
0.00000       27
174.17475      1
8076.81050     1
1.09395        1
18.71595       1
              ..
665.19730      1
1516.36010     1
741.86205      1
183.29890      1
7130.50700     1
Name: generation_gwh_2015, Length: 396, dtype: int64
----------

22. generation_gwh_2016: 403
0.00000       30
8470.57000     2
1511.00000     2
250.97100      1
7.31325        1
              ..
433.84800      1
283.74811      1
259.94375      1
403.96000      1
307.87290      1
Name: generation_gwh_2016, Length: 403, dtype: int64
----------

23. generation_gwh_2017: 408
0.00000       32
170.08530      2
9271.00000     1
59.43135       1
549.86930      1
              ..
214.48220      1
272.73945      1
2887.00000     1
12.73600       1
158.73235      1
Name: generation_gwh_2017, Length: 408, dtype: int64
----------

24. generation_gwh_2018: 410
0.00000       39
100.85320      1
805.48235      1
7179.00000     1
6915.39000     1
              ..
980.25410      1
33.88970       1
6474.61425     1
347.34455      1
192.01510      1
Name: generation_gwh_2018, Length: 410, dtype: int64
----------
```

```
25. generation_gwh_2019: 0
Series([], Name: generation_gwh_2019, dtype: int64)
----------

26. generation_data_source: 1
Central Electricity Authority    449
Name: generation_data_source, dtype: int64
----------

27. estimated_generation_gwh: 0
Series([], Name: estimated_generation_gwh, dtype: int64)
----------
```

In [327…

```
df.describe()
```

Out[327…

	capacity_mw	latitude	longitude	other_fuel3	commissioning_year	wepp_id	year_of_capacity_data	generation_gwh_2013	generation
count	907.000000	861.000000	861.000000	0.0	527.000000	0.0	519.0	0.0	
mean	326.223755	21.197918	77.464907	NaN	1997.091082	NaN	2019.0	NaN	2
std	590.085456	6.239612	4.939316	NaN	17.082868	NaN	0.0	NaN	4
min	0.000000	8.168900	68.644700	NaN	1927.000000	NaN	2019.0	NaN	
25%	16.725000	16.773900	74.256200	NaN	1988.000000	NaN	2019.0	NaN	
50%	59.200000	21.780000	76.719500	NaN	2001.000000	NaN	2019.0	NaN	
75%	385.250000	25.512400	79.440800	NaN	2012.000000	NaN	2019.0	NaN	3
max	4760.000000	34.649000	95.408000	NaN	2018.000000	NaN	2019.0	NaN	28

In [328…

```
##exploring the continuous variables
```

In [329…

```
cont_data = df.select_dtypes(exclude = ['object'] )
cont_data
```

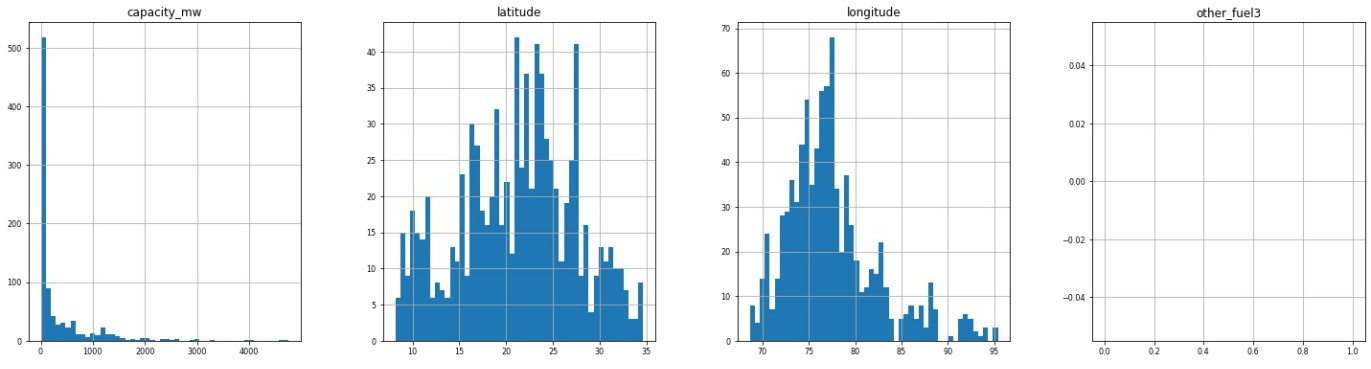
Out[329…

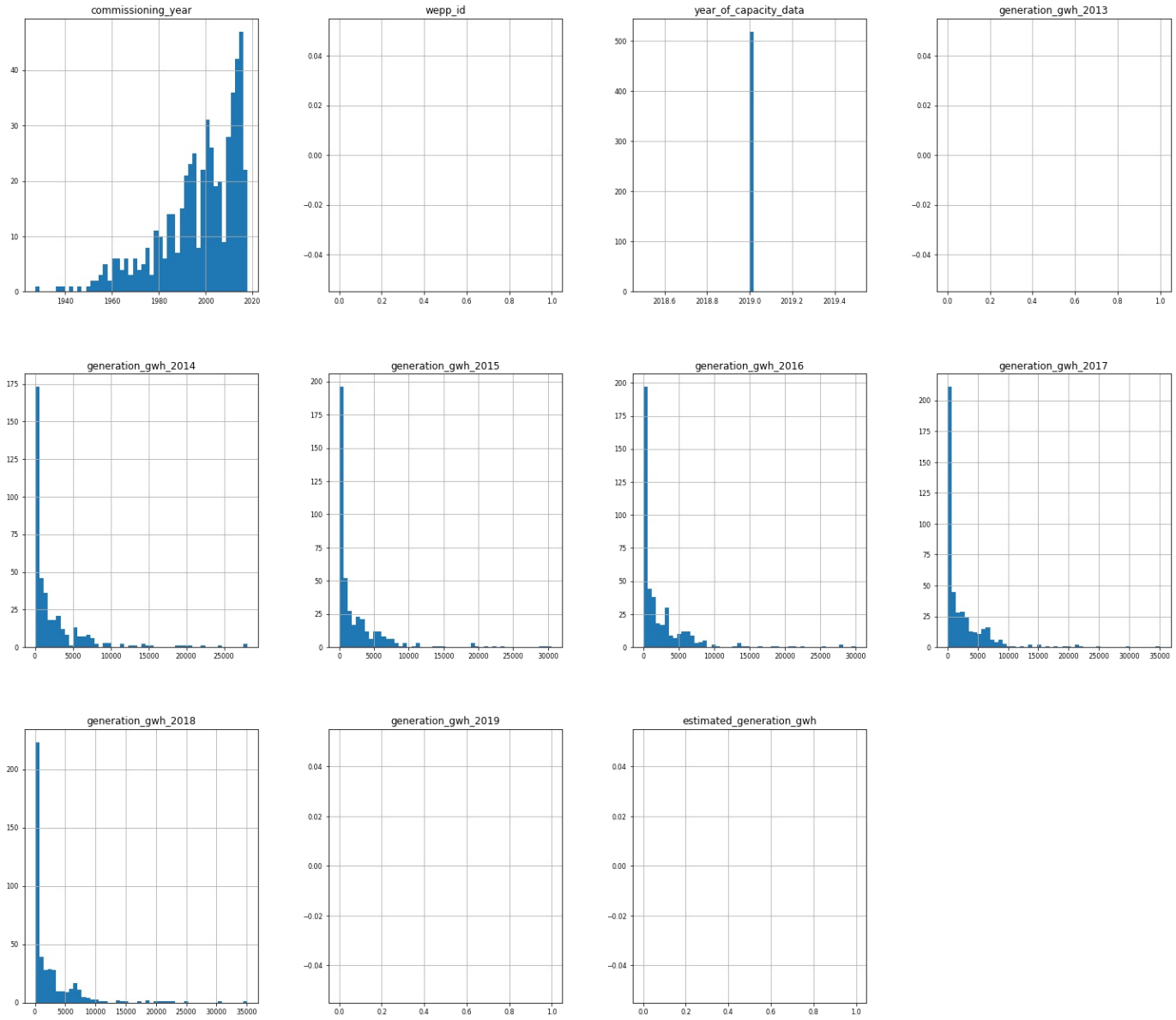
	capacity_mw	latitude	longitude	other_fuel3	commissioning_year	wepp_id	year_of_capacity_data	generation_gwh_2013	generation_gwh_
0	2.5	28.1839	73.2407	NaN	2011.0	NaN	NaN	NaN	
1	98.0	24.7663	74.6090	NaN	NaN	NaN	NaN	NaN	
2	39.2	21.9038	69.3732	NaN	NaN	NaN	NaN	NaN	
3	135.0	23.8712	91.3602	NaN	2004.0	NaN	2019.0	NaN	617.78
4	1800.0	21.9603	82.4091	NaN	2015.0	NaN	2019.0	NaN	3035.55
...	...	...	...	...	...	...	...	...	
902	1600.0	16.2949	77.3568	NaN	2016.0	NaN	2019.0	NaN	
903	3.0	12.8932	78.1654	NaN	NaN	NaN	NaN	NaN	
904	25.5	15.2758	75.5811	NaN	NaN	NaN	NaN	NaN	
905	80.0	24.3500	73.7477	NaN	NaN	NaN	NaN	NaN	
906	16.5	9.9344	77.4768	NaN	NaN	NaN	NaN	NaN	

907 rows × 15 columns

In [330…

```
cont_data.hist(figsize = (25, 30), bins = 50, xlabelsize = 8, ylabelsize = 8)
plt.show()
```





In [331…
```
cont_data.isnull().sum()
```

Out[331…
```
capacity_mw                0
latitude                  46
longitude                 46
other_fuel3              907
commissioning_year       380
wepp_id                  907
year_of_capacity_data    388
generation_gwh_2013      907
generation_gwh_2014      509
generation_gwh_2015      485
generation_gwh_2016      473
generation_gwh_2017      467
generation_gwh_2018      459
generation_gwh_2019      907
estimated_generation_gwh 907
dtype: int64
```

In [332…
```
## dropping the unnecessary columns
```

In [333…
```
##
```

In [334…
```
c=cont_data.drop(['latitude','longitude','other_fuel3','commissioning_year','wepp_id','generation_gwh_2019','esti
c
```

Out[334…

capacity_mw	generation_gwh_2014	generation_gwh_2015	generation_gwh_2016	generation_gwh_2017	generation_gwh_2018

0	2.5	NaN	NaN	NaN	NaN	NaN
1	98.0	NaN	NaN	NaN	NaN	NaN
2	39.2	NaN	NaN	NaN	NaN	NaN
3	135.0	617.789264	843.747000	886.004428	663.774500	626.239128
4	1800.0	3035.550000	5916.370000	6243.000000	5385.579736	7279.000000
...	...	...	...	...	...	...
902	1600.0	NaN	0.994875	233.596650	865.400000	686.500000
903	3.0	NaN	NaN	NaN	NaN	NaN
904	25.5	NaN	NaN	NaN	NaN	NaN
905	80.0	NaN	NaN	NaN	NaN	NaN
906	16.5	NaN	NaN	NaN	NaN	NaN

907 rows × 6 columns

In [335…
```
c.isnull().sum()
```

Out[335…
```
capacity_mw              0
generation_gwh_2014    509
generation_gwh_2015    485
generation_gwh_2016    473
generation_gwh_2017    467
generation_gwh_2018    459
dtype: int64
```

In [336…
```
c['generation_gwh_2014'].fillna(c['generation_gwh_2014'].mean(),inplace=True)
```

In [337…
```
c
```

Out[337…

	capacity_mw	generation_gwh_2014	generation_gwh_2015	generation_gwh_2016	generation_gwh_2017	generation_gwh_2018
0	2.5	2431.823590	NaN	NaN	NaN	NaN
1	98.0	2431.823590	NaN	NaN	NaN	NaN
2	39.2	2431.823590	NaN	NaN	NaN	NaN
3	135.0	617.789264	843.747000	886.004428	663.774500	626.239128
4	1800.0	3035.550000	5916.370000	6243.000000	5385.579736	7279.000000
...	...	...	...	...	...	...
902	1600.0	2431.823590	0.994875	233.596650	865.400000	686.500000
903	3.0	2431.823590	NaN	NaN	NaN	NaN
904	25.5	2431.823590	NaN	NaN	NaN	NaN
905	80.0	2431.823590	NaN	NaN	NaN	NaN
906	16.5	2431.823590	NaN	NaN	NaN	NaN

907 rows × 6 columns

In [338…
```
c['generation_gwh_2015'].fillna(c['generation_gwh_2015'].mean(),inplace=True)
```

In [339…
```
c
```

Out[339…

	capacity_mw	generation_gwh_2014	generation_gwh_2015	generation_gwh_2016	generation_gwh_2017	generation_gwh_2018
0	2.5	2431.823590	2428.226946	NaN	NaN	NaN
1	98.0	2431.823590	2428.226946	NaN	NaN	NaN
2	39.2	2431.823590	2428.226946	NaN	NaN	NaN
3	135.0	617.789264	843.747000	886.004428	663.774500	626.239128
4	1800.0	3035.550000	5916.370000	6243.000000	5385.579736	7279.000000
...	...	...	...	...	...	...
902	1600.0	2431.823590	0.994875	233.596650	865.400000	686.500000

903	3.0	2431.823590	2428.226946	NaN	NaN	NaN
904	25.5	2431.823590	2428.226946	NaN	NaN	NaN
905	80.0	2431.823590	2428.226946	NaN	NaN	NaN
906	16.5	2431.823590	2428.226946	NaN	NaN	NaN

907 rows × 6 columns

In [340…
```
c['generation_gwh_2016'].fillna(c['generation_gwh_2016'].mean(),inplace=True)
```

In [341…
```
c
```

Out[341…

	capacity_mw	generation_gwh_2014	generation_gwh_2015	generation_gwh_2016	generation_gwh_2017	generation_gwh_2018
0	2.5	2431.823590	2428.226946	2467.936859	NaN	NaN
1	98.0	2431.823590	2428.226946	2467.936859	NaN	NaN
2	39.2	2431.823590	2428.226946	2467.936859	NaN	NaN
3	135.0	617.789264	843.747000	886.004428	663.774500	626.239128
4	1800.0	3035.550000	5916.370000	6243.000000	5385.579736	7279.000000
...	...	...	...	...	...	...
902	1600.0	2431.823590	0.994875	233.596650	865.400000	686.500000
903	3.0	2431.823590	2428.226946	2467.936859	NaN	NaN
904	25.5	2431.823590	2428.226946	2467.936859	NaN	NaN
905	80.0	2431.823590	2428.226946	2467.936859	NaN	NaN
906	16.5	2431.823590	2428.226946	2467.936859	NaN	NaN

907 rows × 6 columns

In [342…
```
c['generation_gwh_2017'].fillna(c['generation_gwh_2017'].mean(),inplace=True)
```

In [343…
```
c
```

Out[343…

	capacity_mw	generation_gwh_2014	generation_gwh_2015	generation_gwh_2016	generation_gwh_2017	generation_gwh_2018
0	2.5	2431.823590	2428.226946	2467.936859	2547.759305	NaN
1	98.0	2431.823590	2428.226946	2467.936859	2547.759305	NaN
2	39.2	2431.823590	2428.226946	2467.936859	2547.759305	NaN
3	135.0	617.789264	843.747000	886.004428	663.774500	626.239128
4	1800.0	3035.550000	5916.370000	6243.000000	5385.579736	7279.000000
...	...	...	...	...	...	...
902	1600.0	2431.823590	0.994875	233.596650	865.400000	686.500000
903	3.0	2431.823590	2428.226946	2467.936859	2547.759305	NaN
904	25.5	2431.823590	2428.226946	2467.936859	2547.759305	NaN
905	80.0	2431.823590	2428.226946	2467.936859	2547.759305	NaN
906	16.5	2431.823590	2428.226946	2467.936859	2547.759305	NaN

907 rows × 6 columns

In [344…
```
c['generation_gwh_2018'].fillna(c['generation_gwh_2018'].mean(),inplace=True)
```

In [345…
```
c
```

Out[345…

	capacity_mw	generation_gwh_2014	generation_gwh_2015	generation_gwh_2016	generation_gwh_2017	generation_gwh_2018
0	2.5	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099
1	98.0	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099
2	39.2	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099
3	135.0	617.789264	843.747000	886.004428	663.774500	626.239128

4	1800.0	3035.550000	5916.370000	6243.000000	5385.579736	7279.000000
...	...	...	...	...	...	...
902	1600.0	2431.823590	0.994875	233.596650	865.400000	686.500000
903	3.0	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099
904	25.5	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099
905	80.0	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099
906	16.5	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099

907 rows × 6 columns

In [346…
```
c.isnull().sum()
```

Out[346…
```
capacity_mw            0
generation_gwh_2014    0
generation_gwh_2015    0
generation_gwh_2016    0
generation_gwh_2017    0
generation_gwh_2018    0
dtype: int64
```

In [347…
```
## looks like there are no null values
```

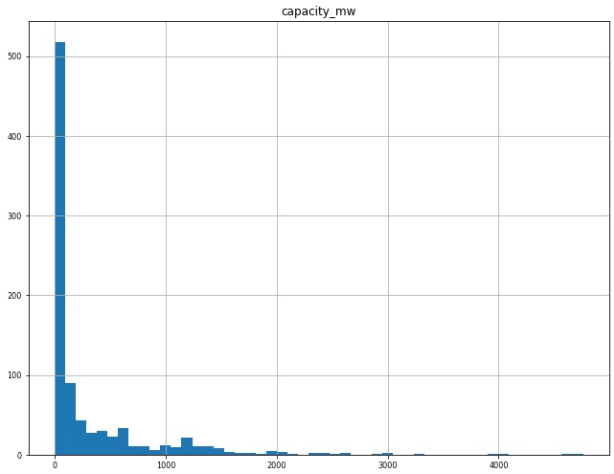
In [348…
```
c.describe()
```

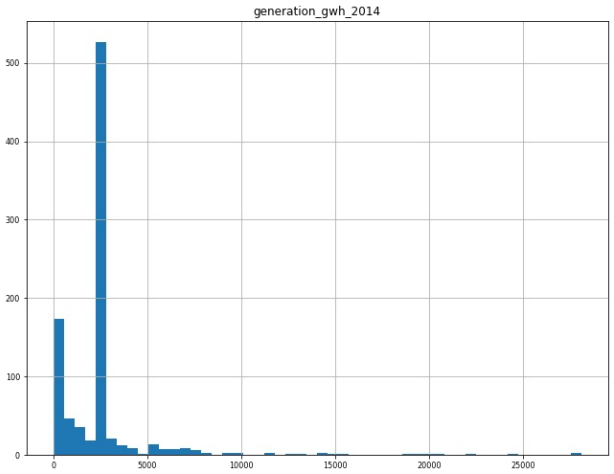
Out[348…

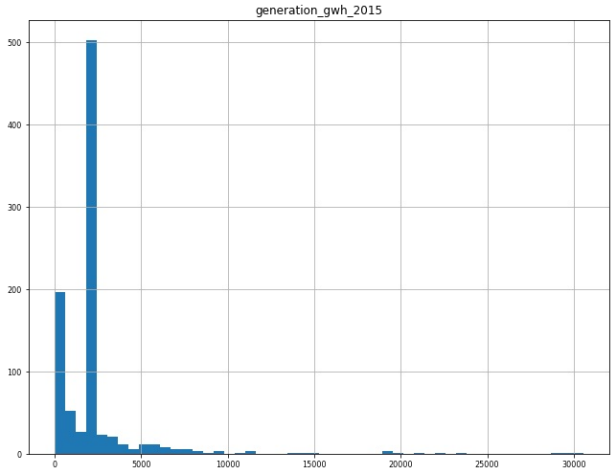
	capacity_mw	generation_gwh_2014	generation_gwh_2015	generation_gwh_2016	generation_gwh_2017	generation_gwh_2018
count	907.000000	907.000000	907.000000	907.000000	907.000000	907.000000
mean	326.223755	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099
std	590.085456	2665.338608	2859.349132	2877.890004	2921.502193	3030.808041
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	16.725000	1211.362750	916.000000	896.500214	882.594850	824.842340
50%	59.200000	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099
75%	385.250000	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099
max	4760.000000	28127.000000	30539.000000	30015.000000	35116.000000	35136.000000

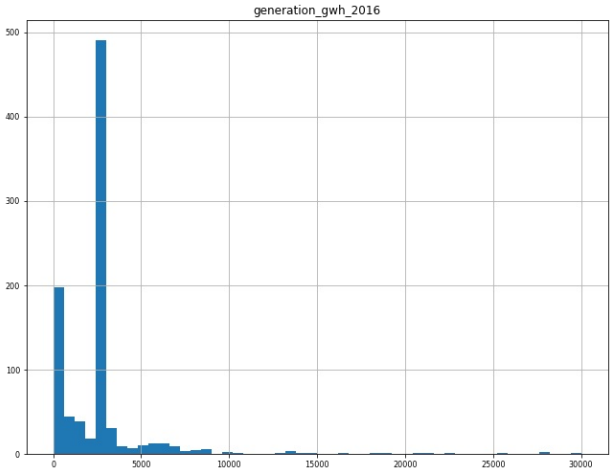
In [349…
```
c.hist(figsize = (25, 30), bins = 50, xlabelsize = 8, ylabelsize = 8)
```

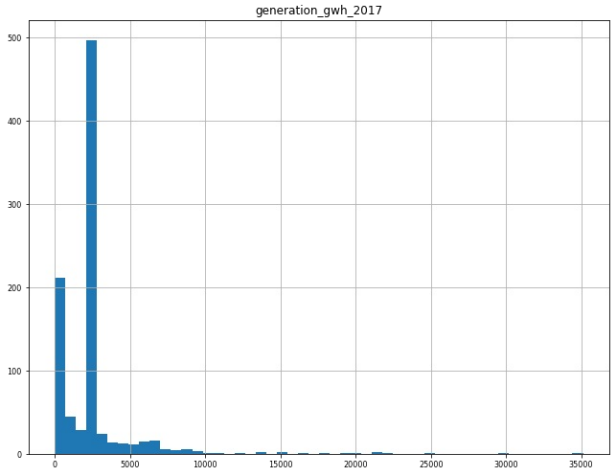
Out[349…
```
array([[<AxesSubplot:title={'center':'capacity_mw'}>,
        <AxesSubplot:title={'center':'generation_gwh_2014'}>],
       [<AxesSubplot:title={'center':'generation_gwh_2015'}>,
        <AxesSubplot:title={'center':'generation_gwh_2016'}>],
       [<AxesSubplot:title={'center':'generation_gwh_2017'}>,
        <AxesSubplot:title={'center':'generation_gwh_2018'}>]],
      dtype=object)
```

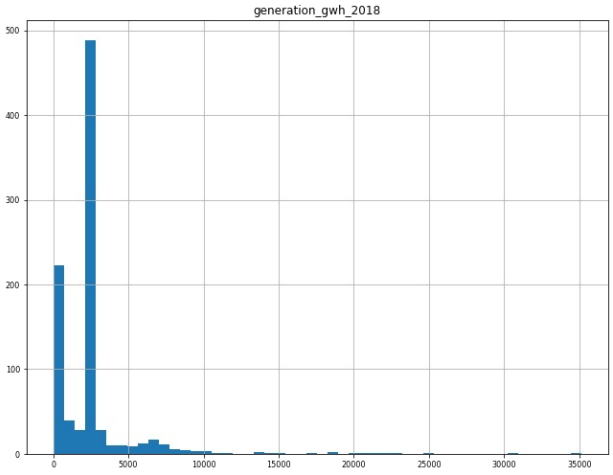






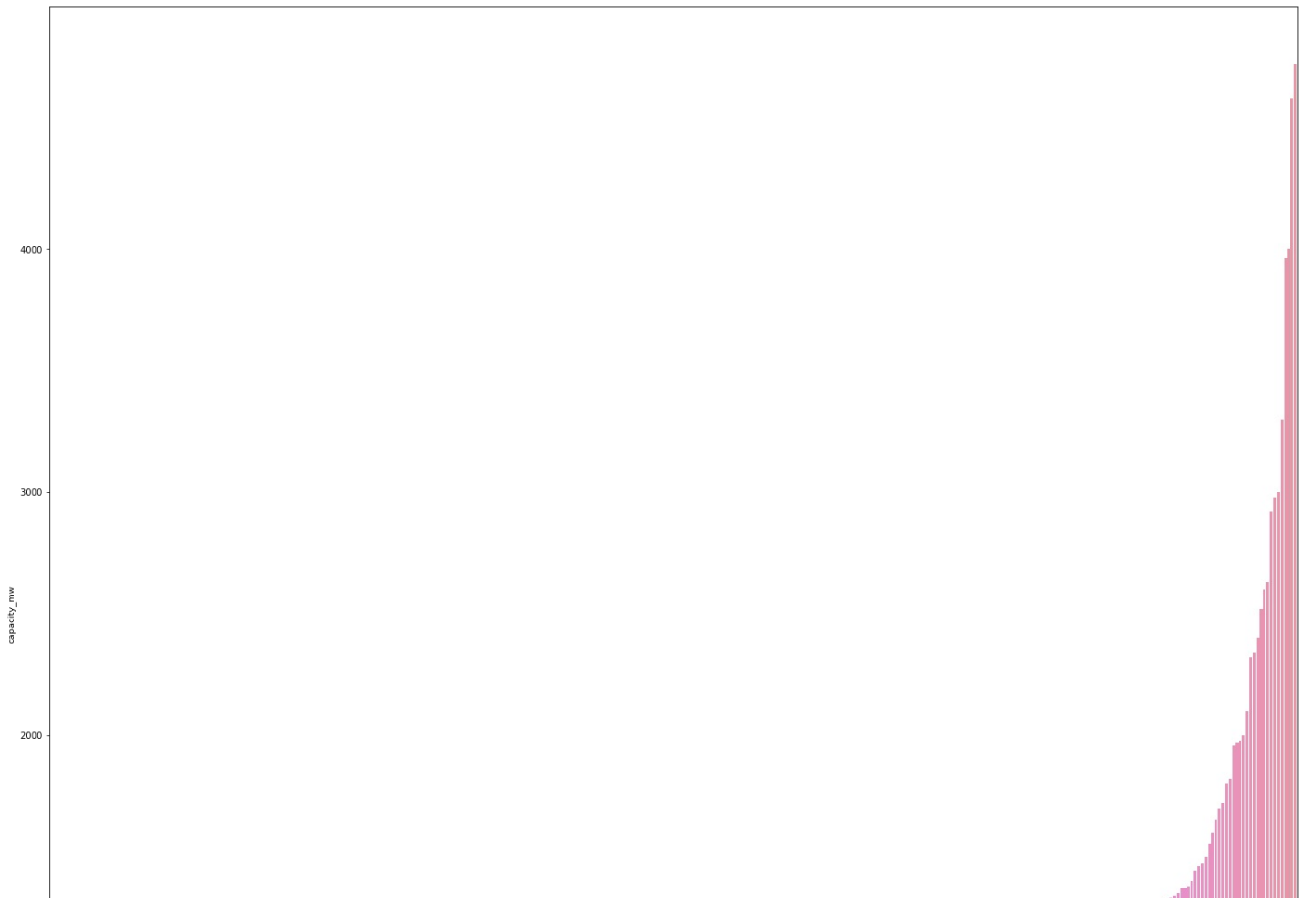


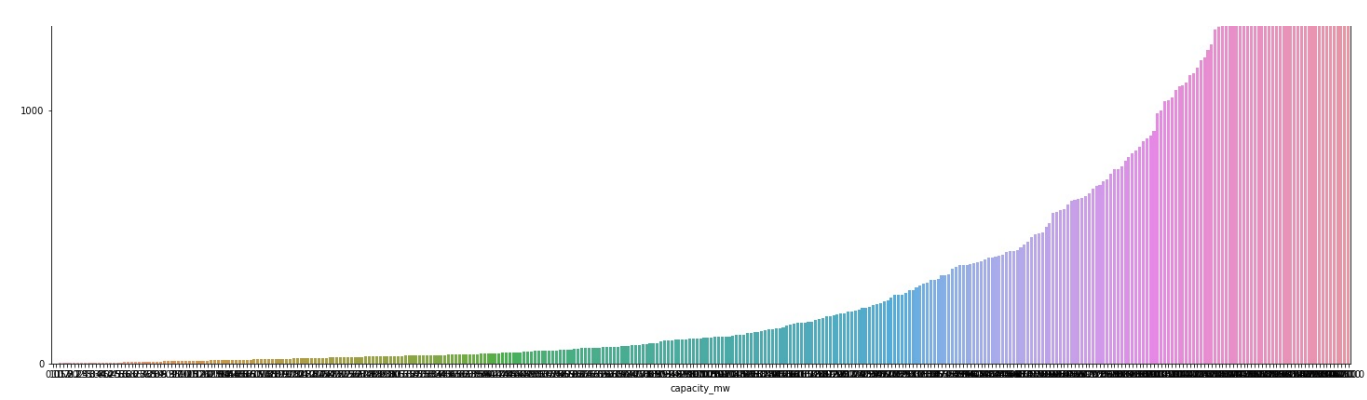


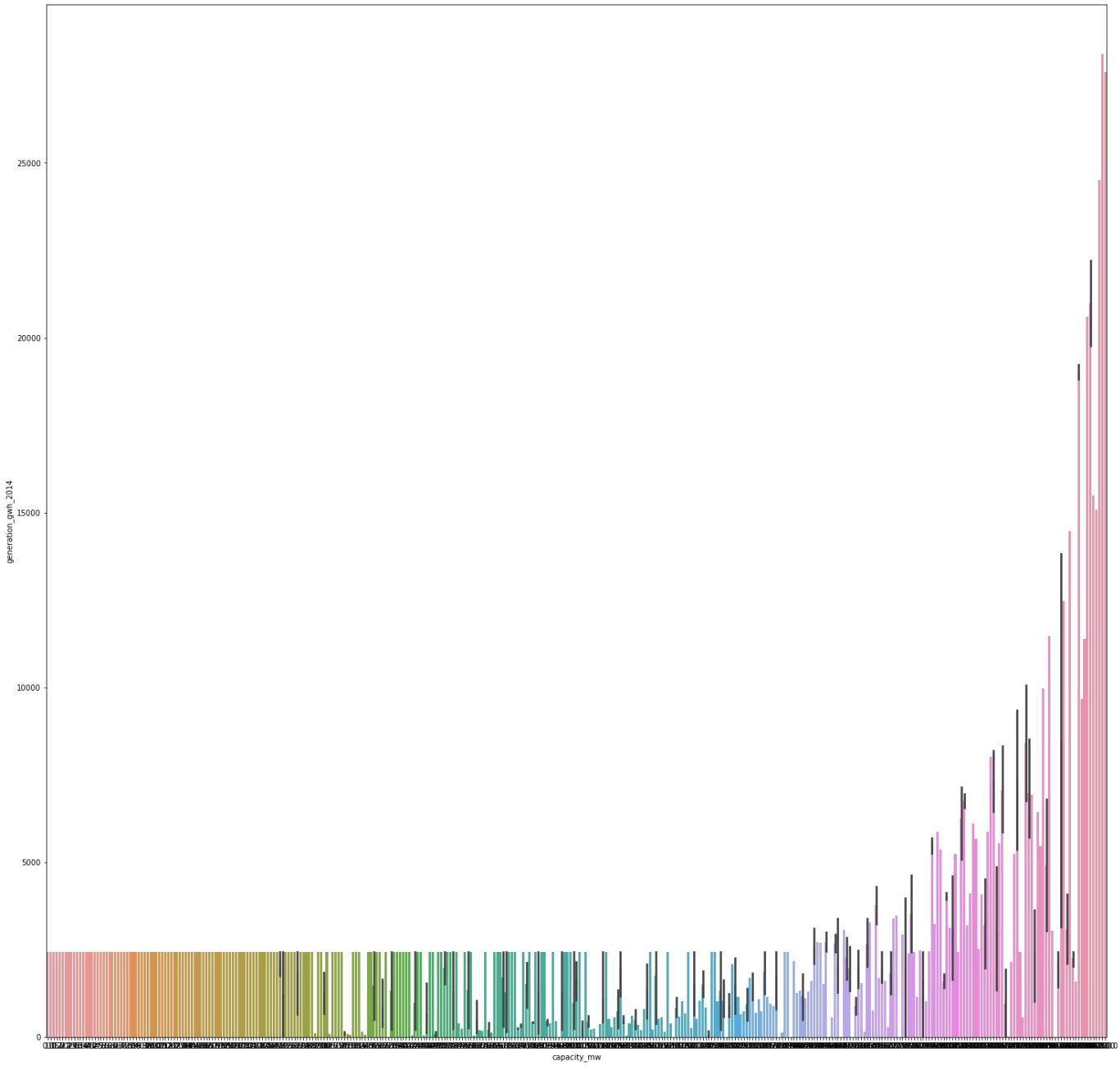


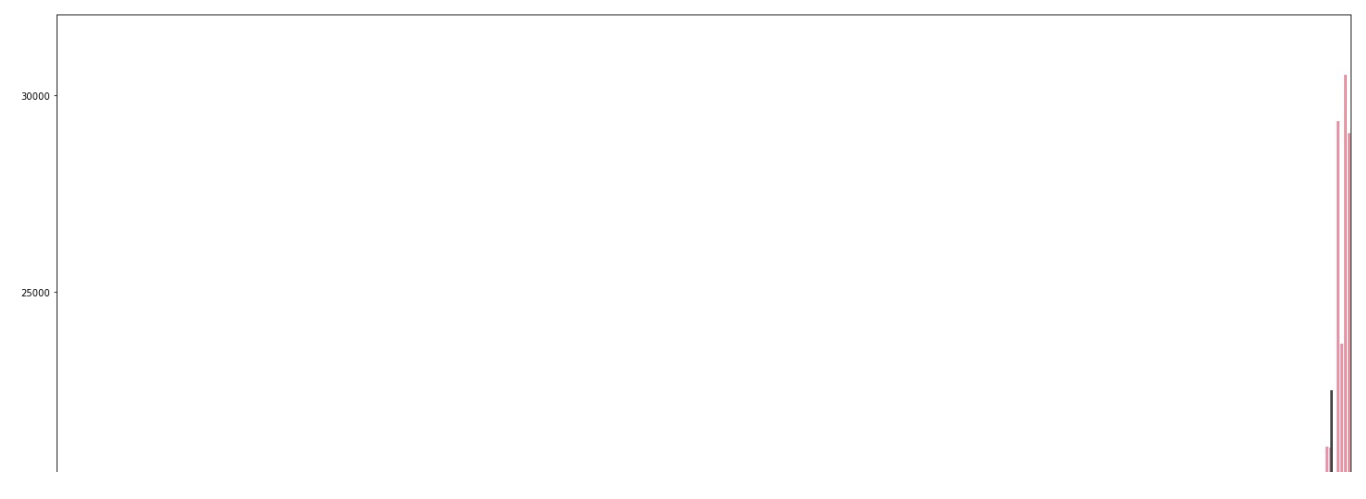
In [350…

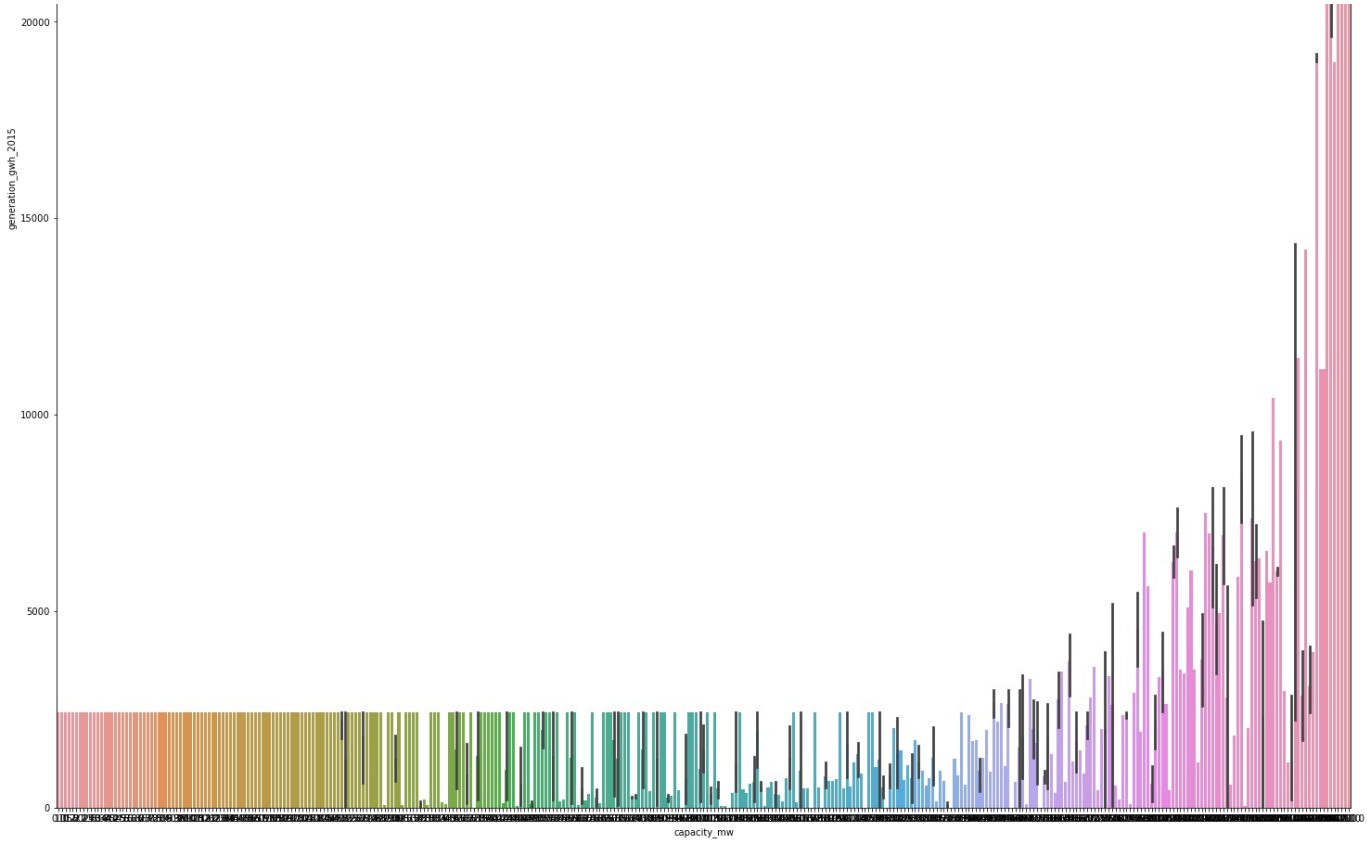
```
for i in c:
    sns.barplot(y = c[i], x = df['capacity_mw'])
    plt.show()
```

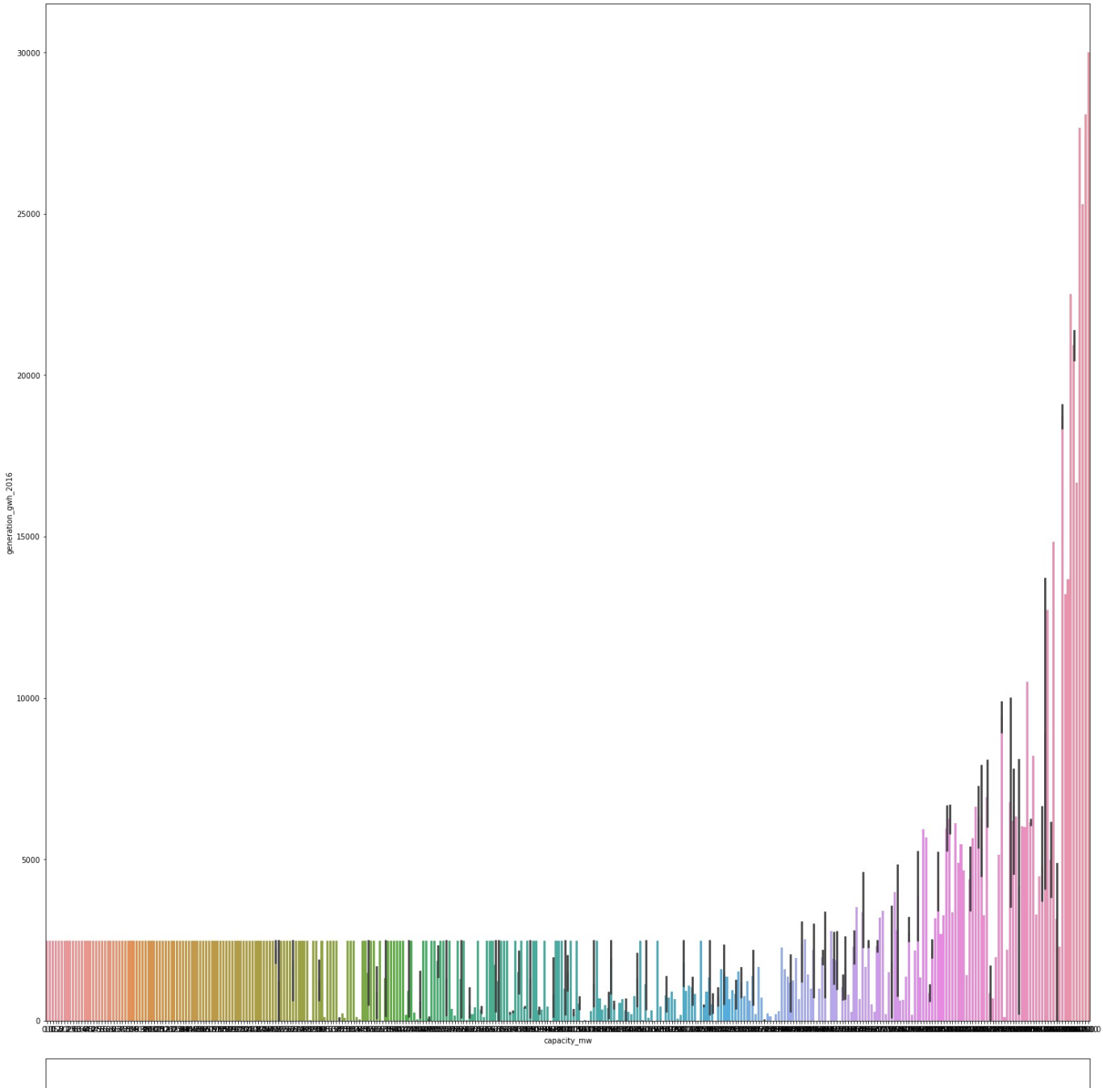


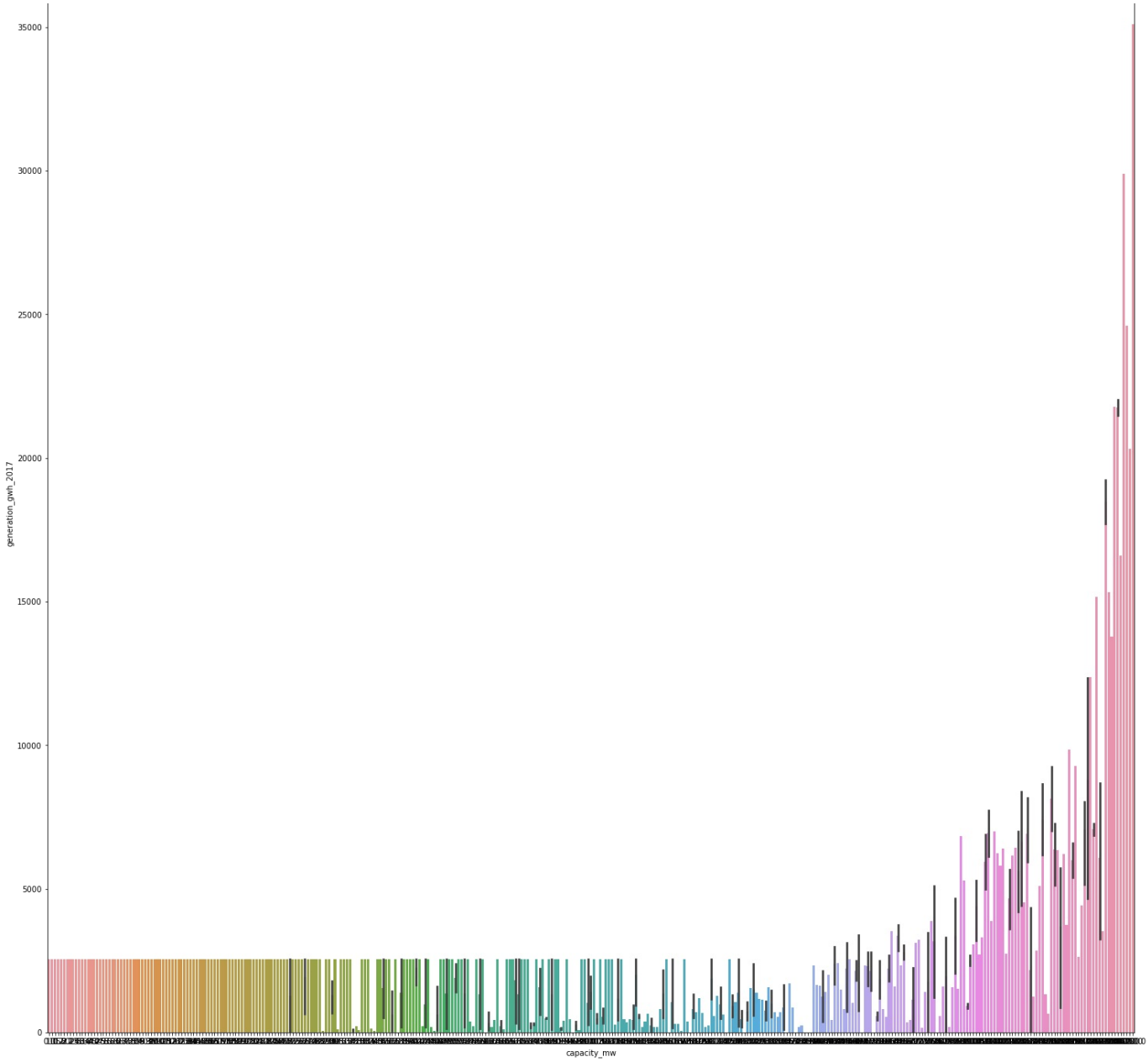

1000
0
capacity_mw

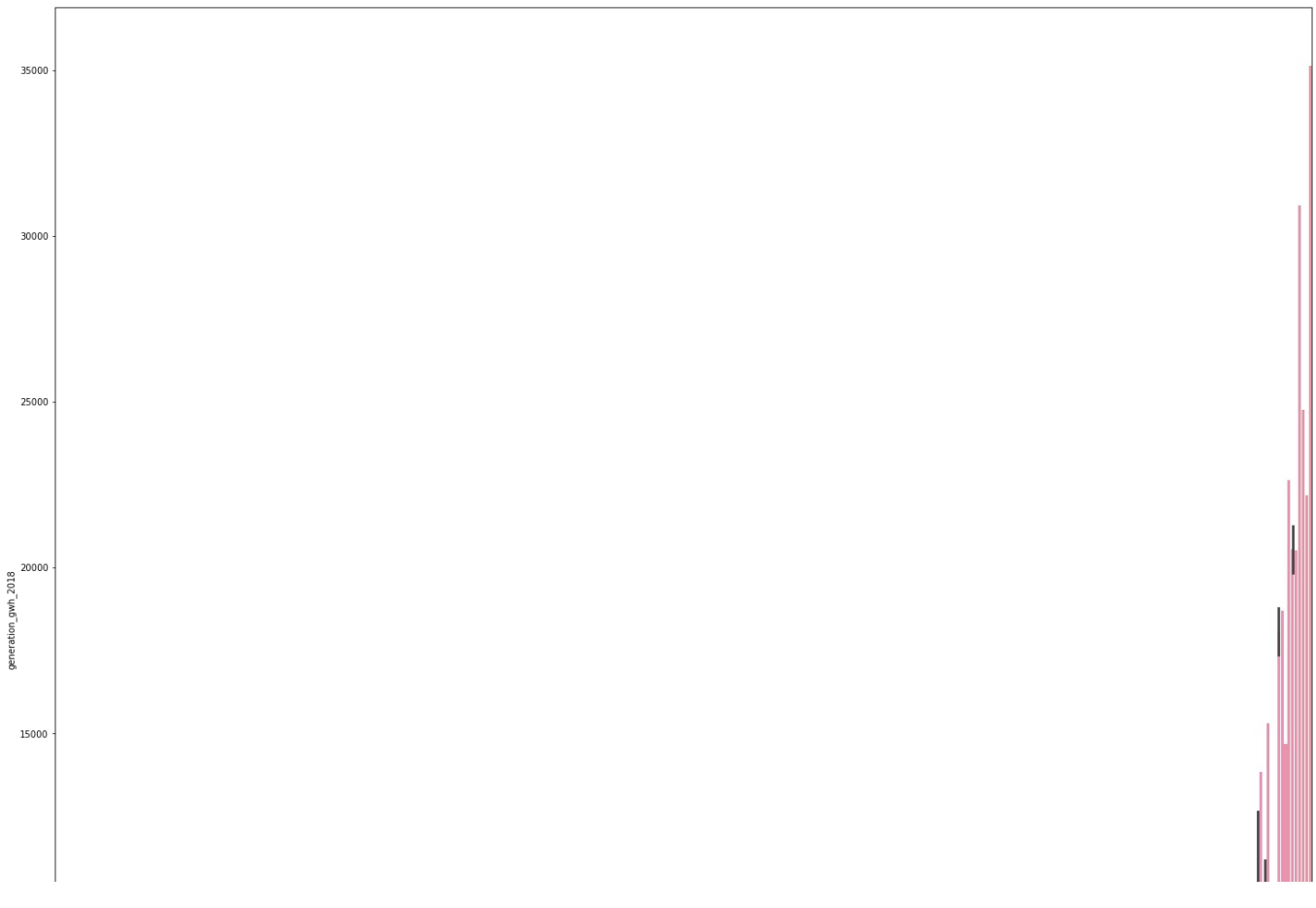

25000
20000
15000
10000
5000
0
generation_gwh_2014
capacity_mw

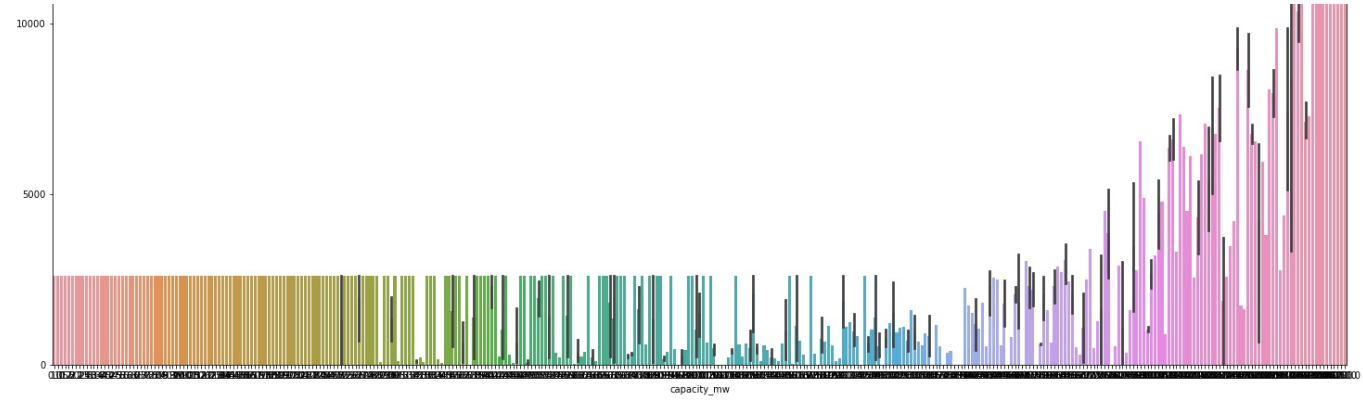

30000
25000


generation_gwh_2015
20000
15000
10000
5000
0
capacity_mw


generation_gwh_2016
30000
25000
20000
15000
10000
5000
0
capacity_mw


35000
30000
25000
20000
15000
10000
5000
0
generation_gwh_2017
capacity_mw


35000
30000
25000
20000
15000
generation_gwh_2018

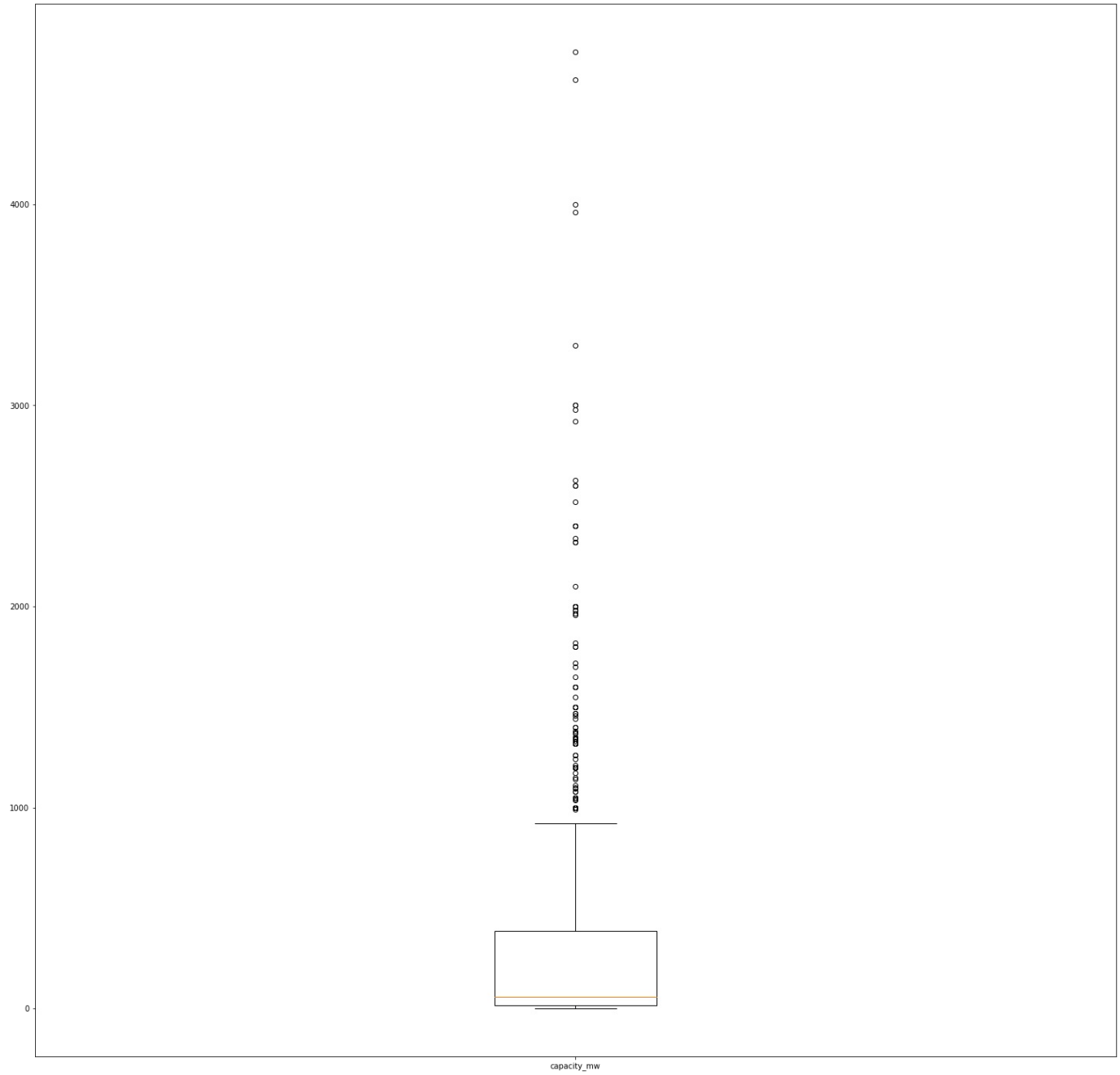


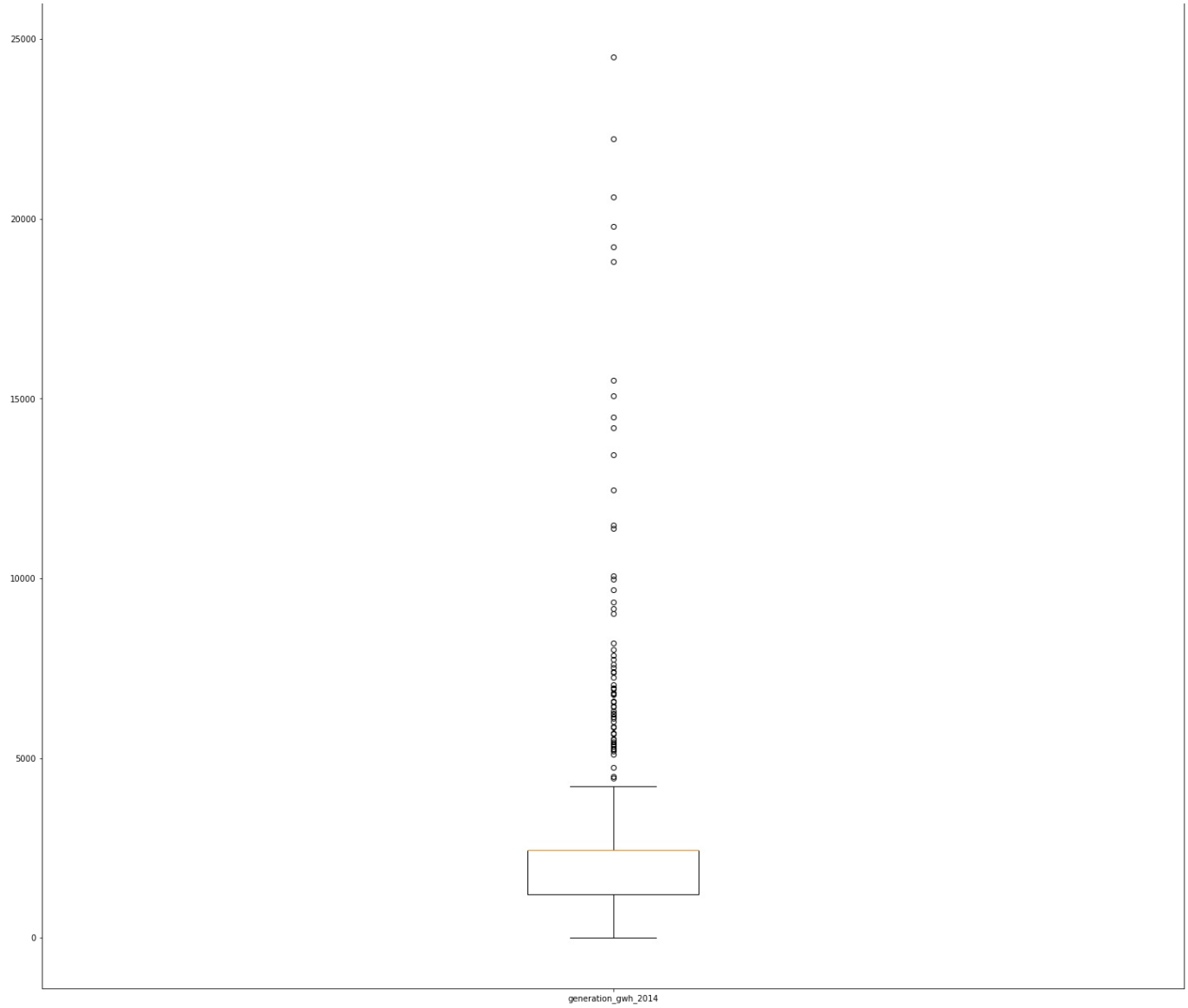
In [351…

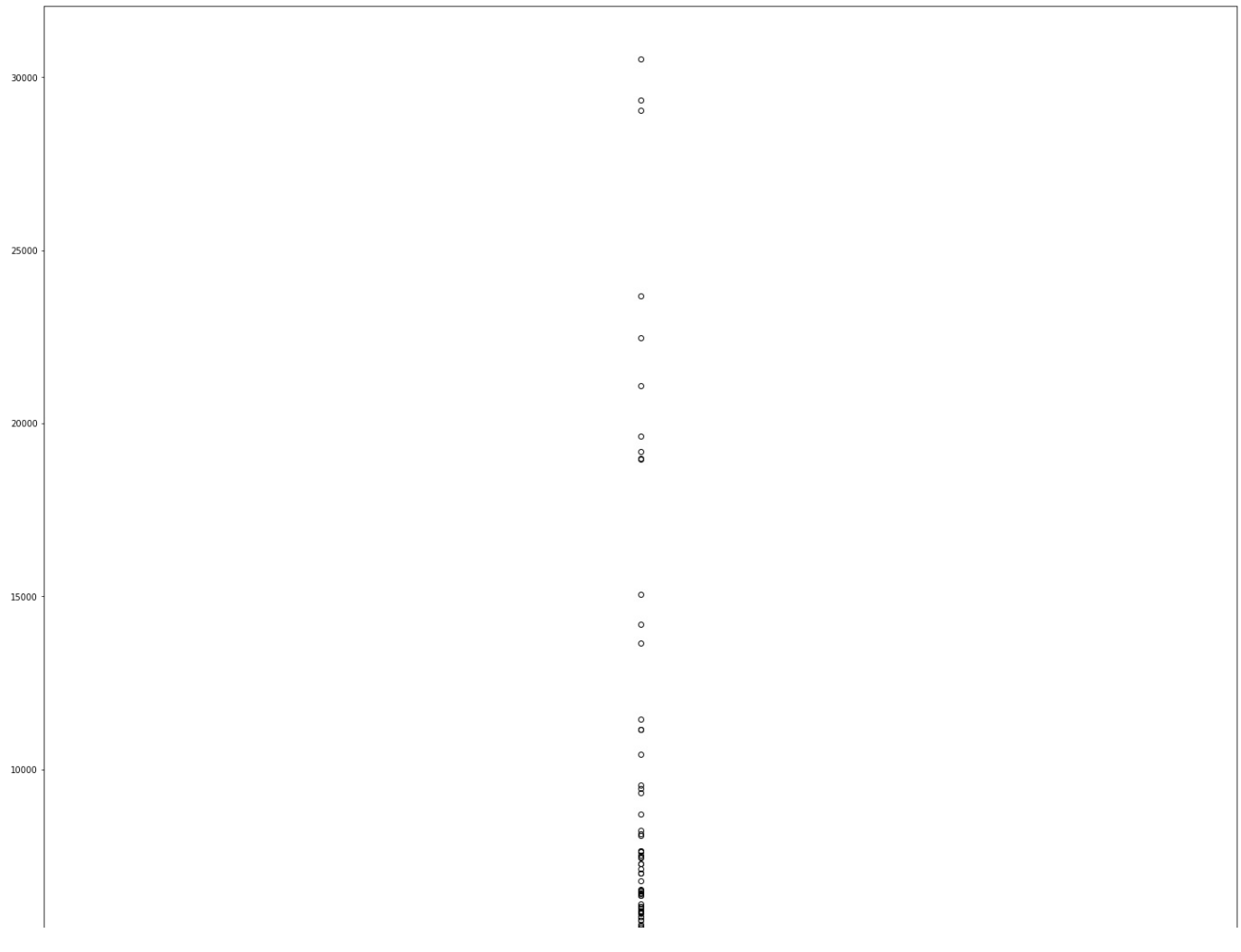
```
## checking for outliers
```

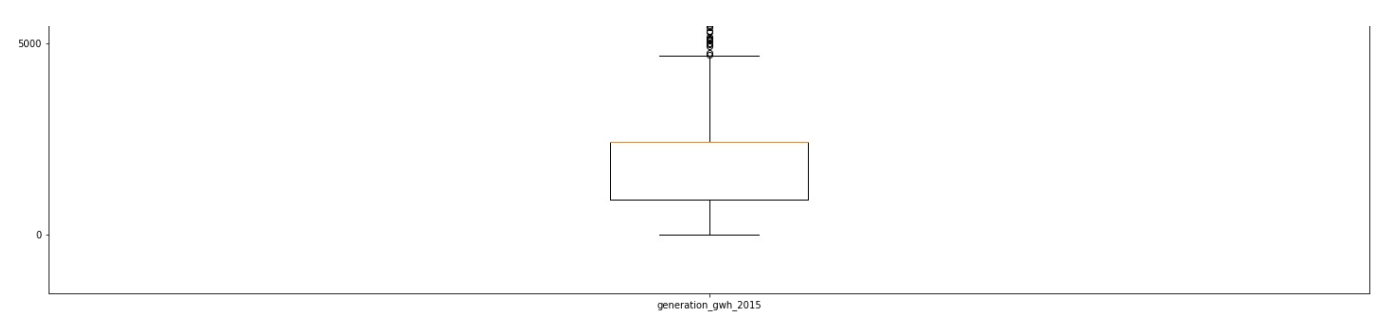
In [352…

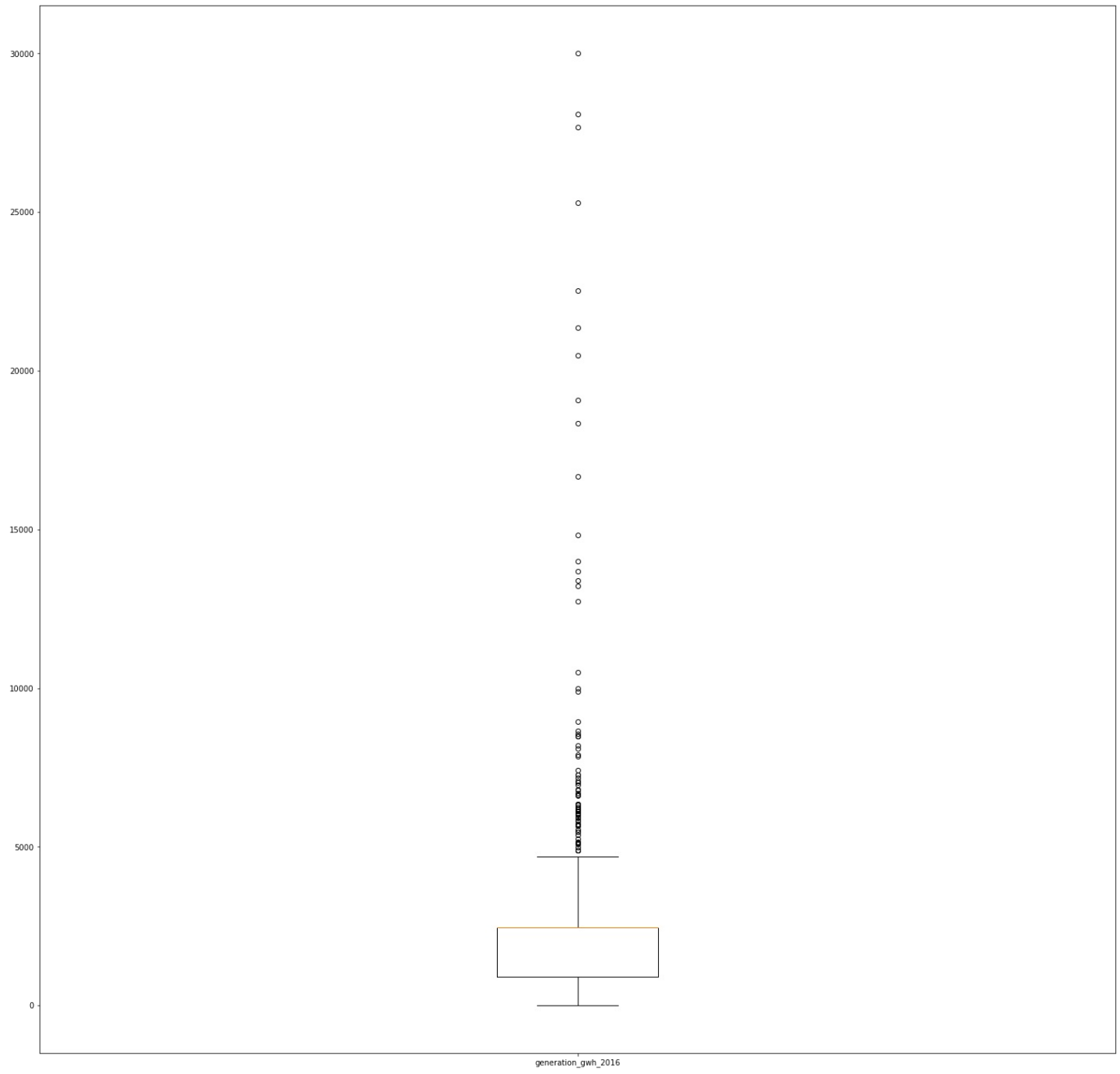
```
for i in c:
    plt.boxplot(c[i], labels = [i])
    plt.show()
```

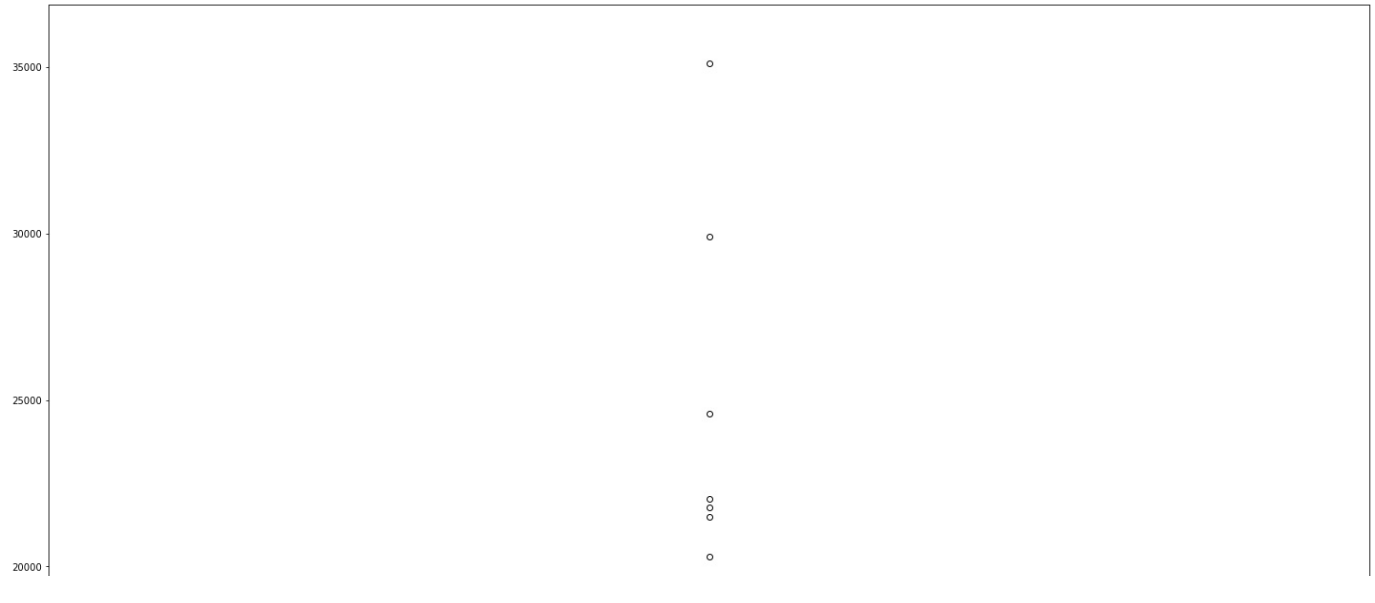


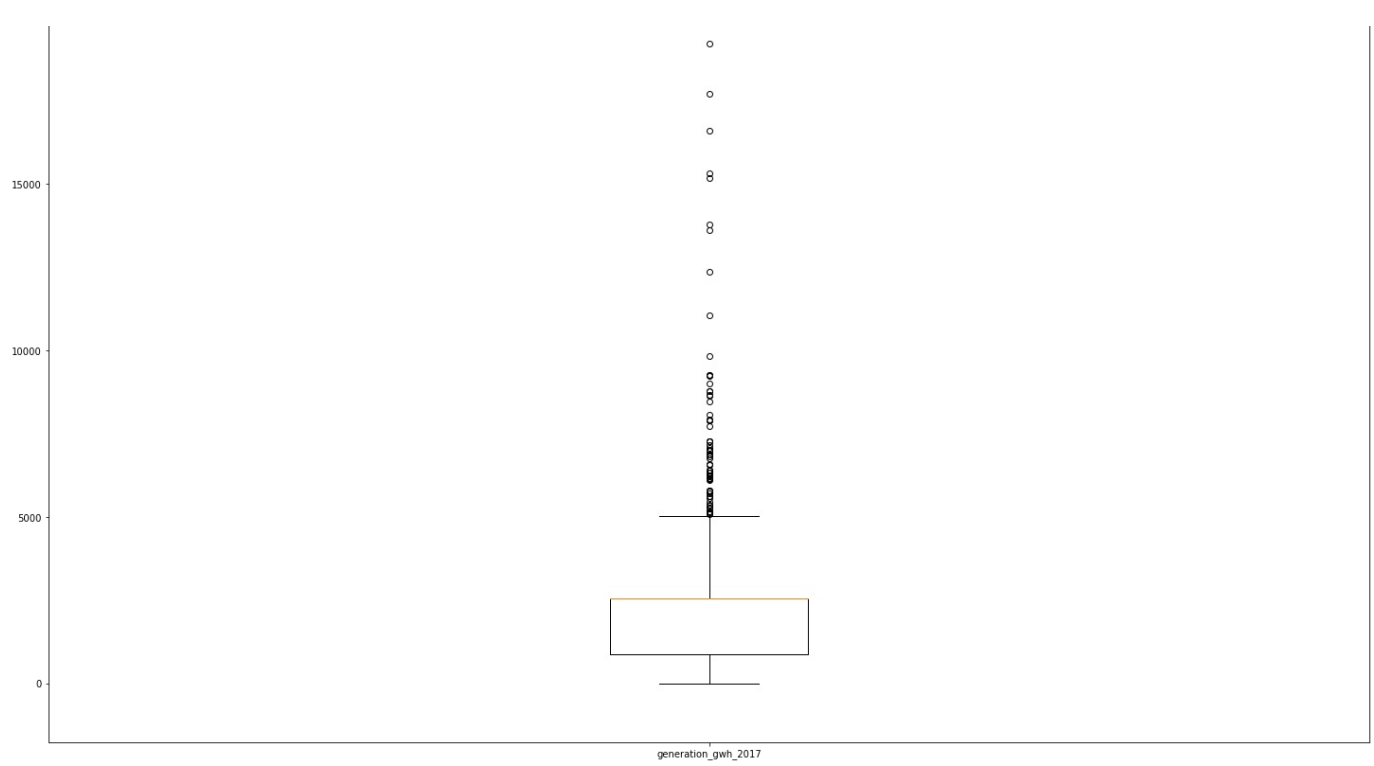

25000
20000
15000
10000
5000
0
generation_gwh_2014

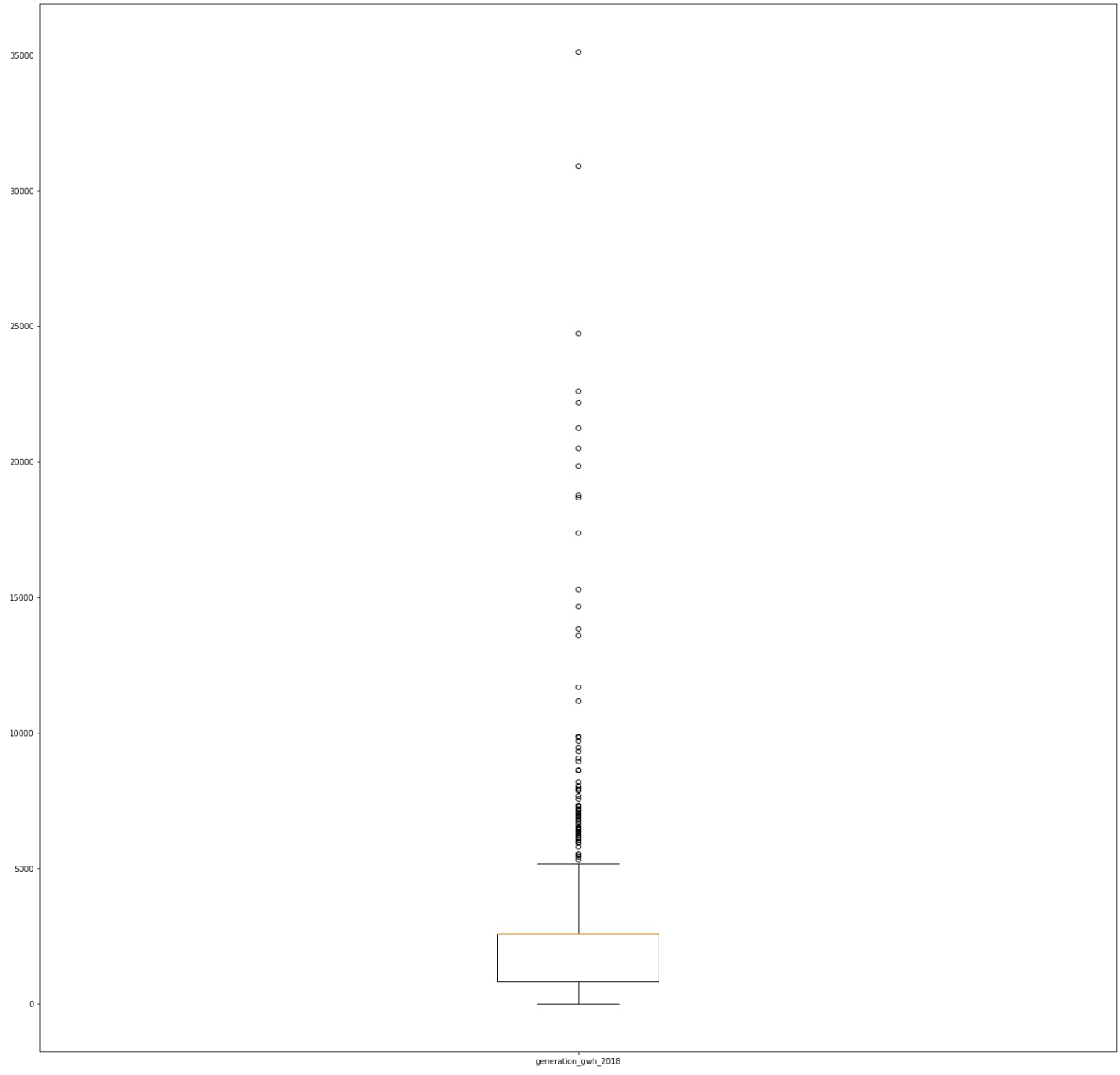

30000
25000
20000
15000
10000


5000
0
generation_gwh_2015


30000
25000
20000
15000
10000
5000
0
generation_gwh_2016


35000
30000
25000
20000


15000
10000
5000
0
generation_gwh_2017


35000
30000
25000
20000
15000
10000
5000
0
generation_gwh_2018

In [353…

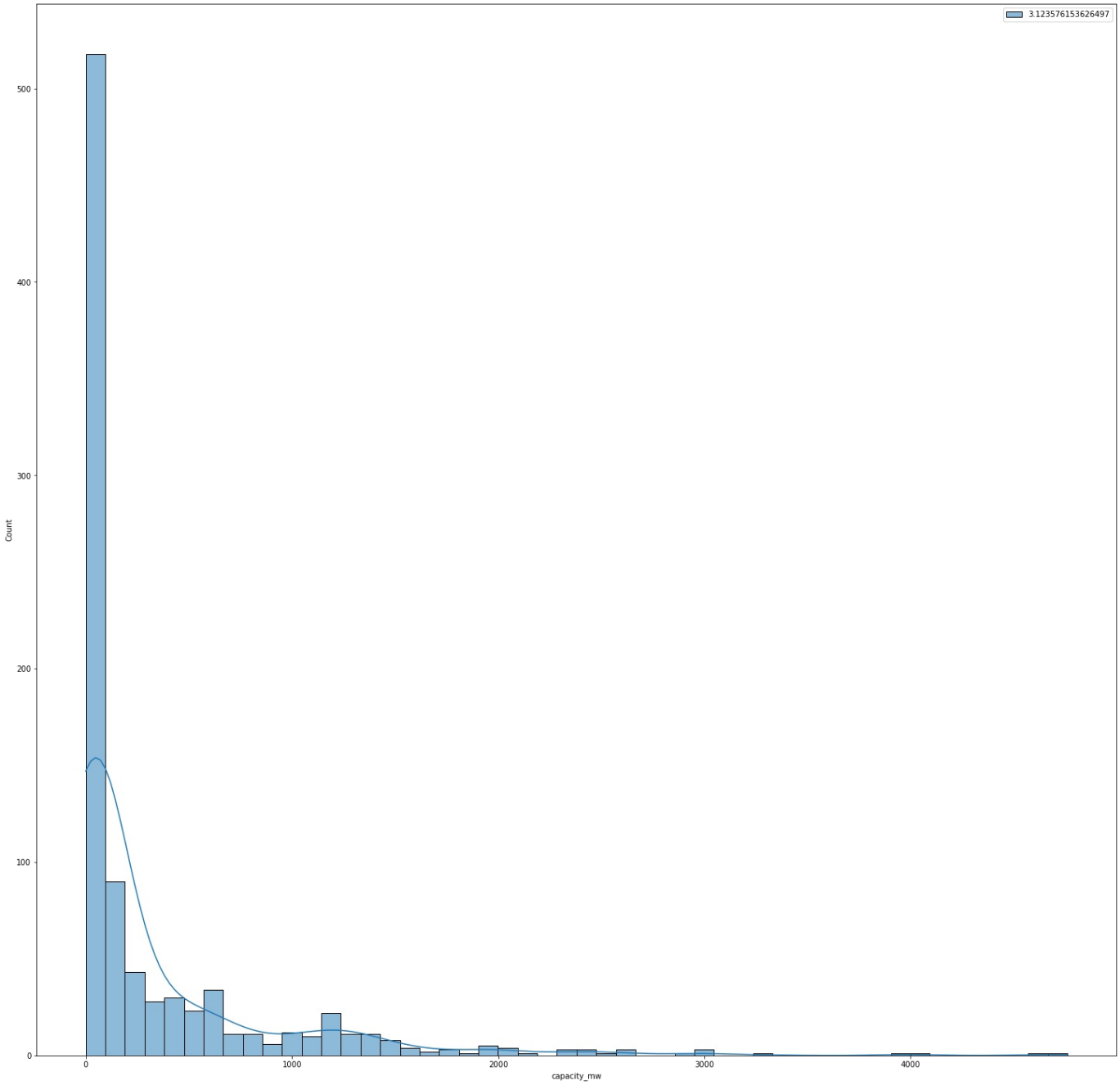
```
## looks like there are lot of outliers
```

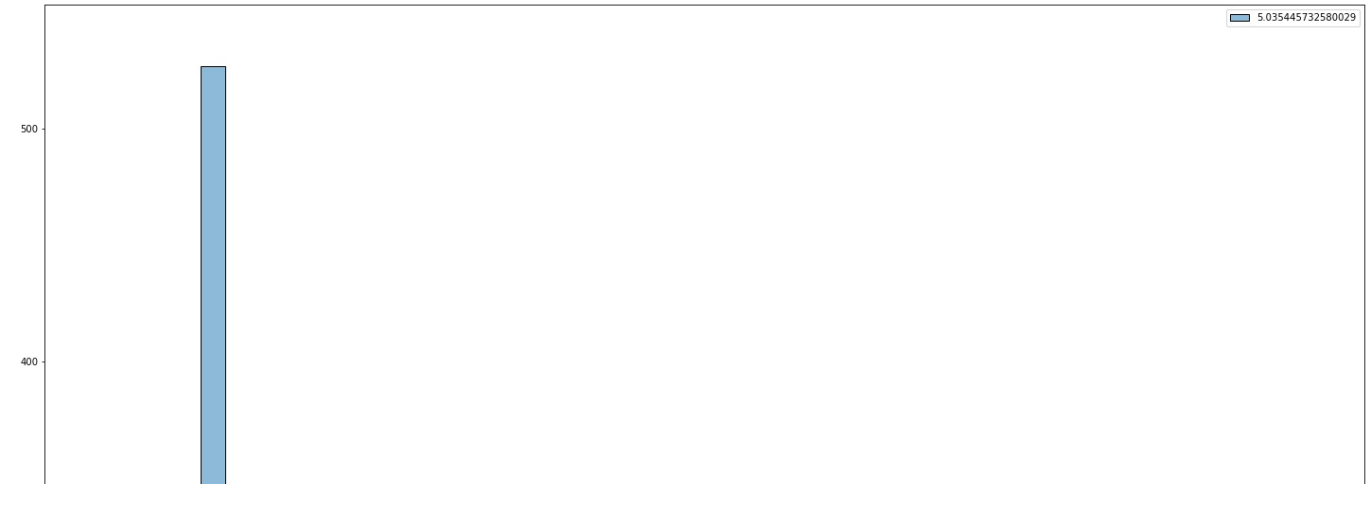
In [354…

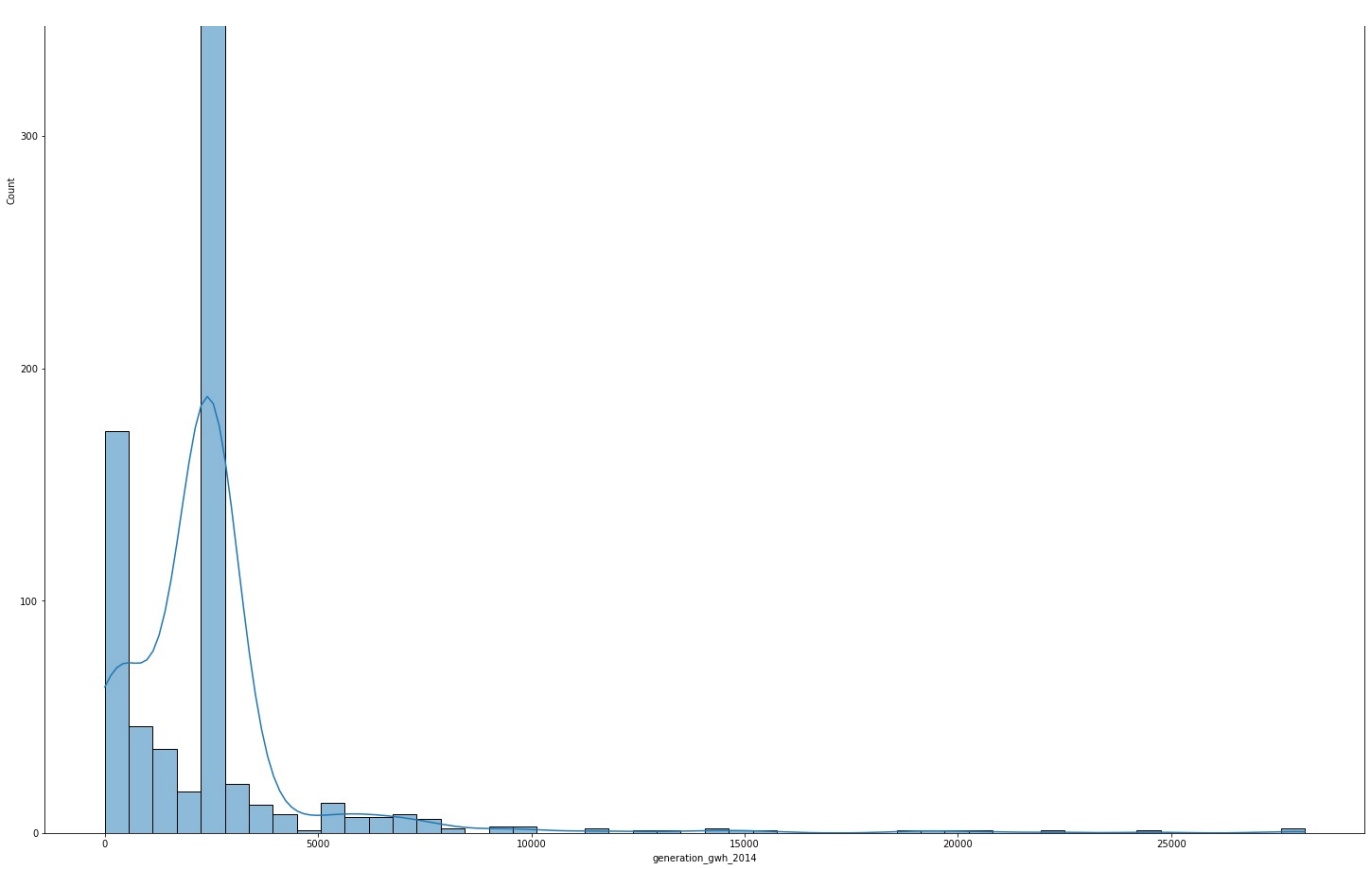
```
a=['capacity_mw','generation_gwh_2014','generation_gwh_2015','generation_gwh_2016','generation_gwh_2017','generat
```

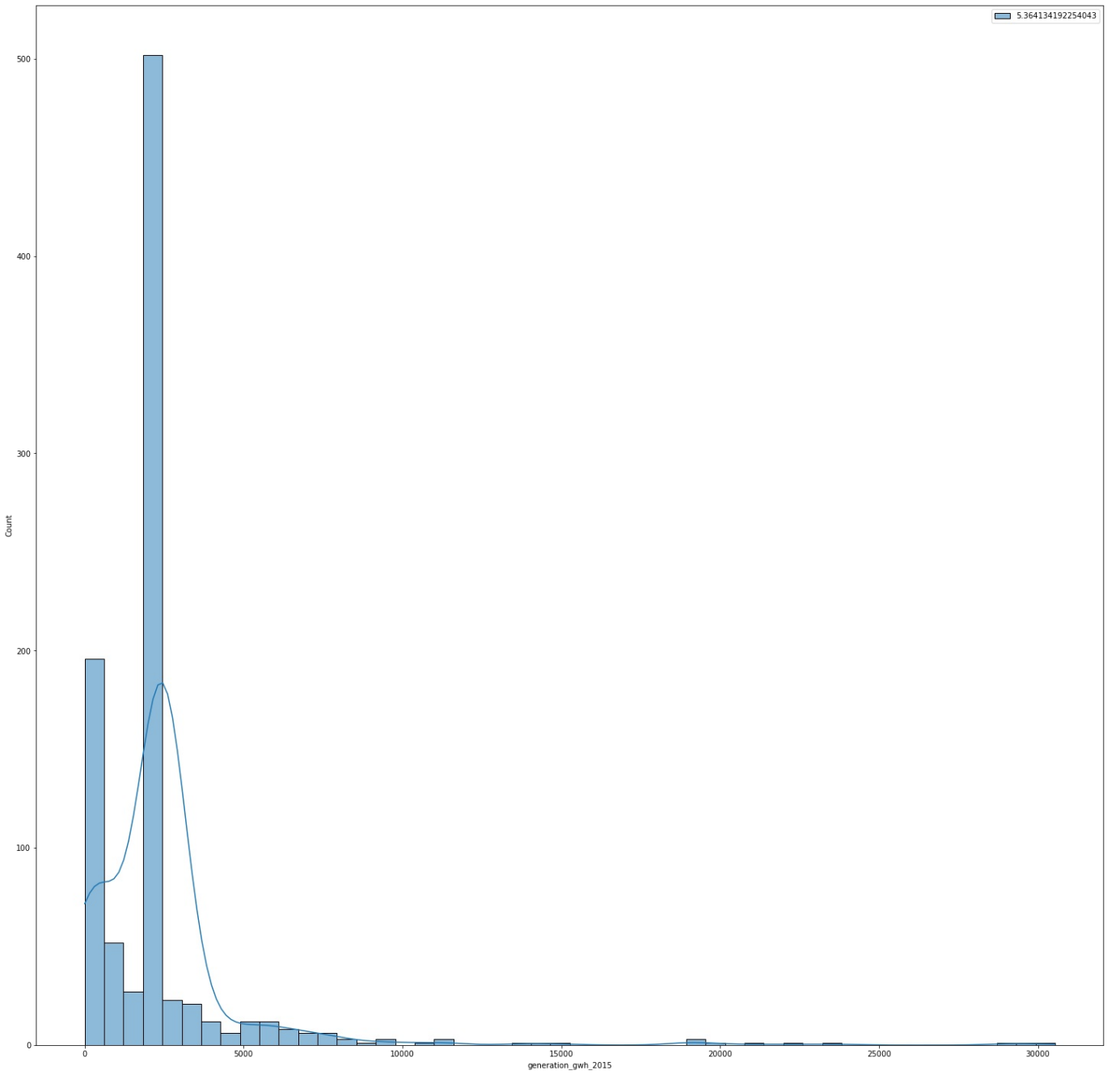
In [355…

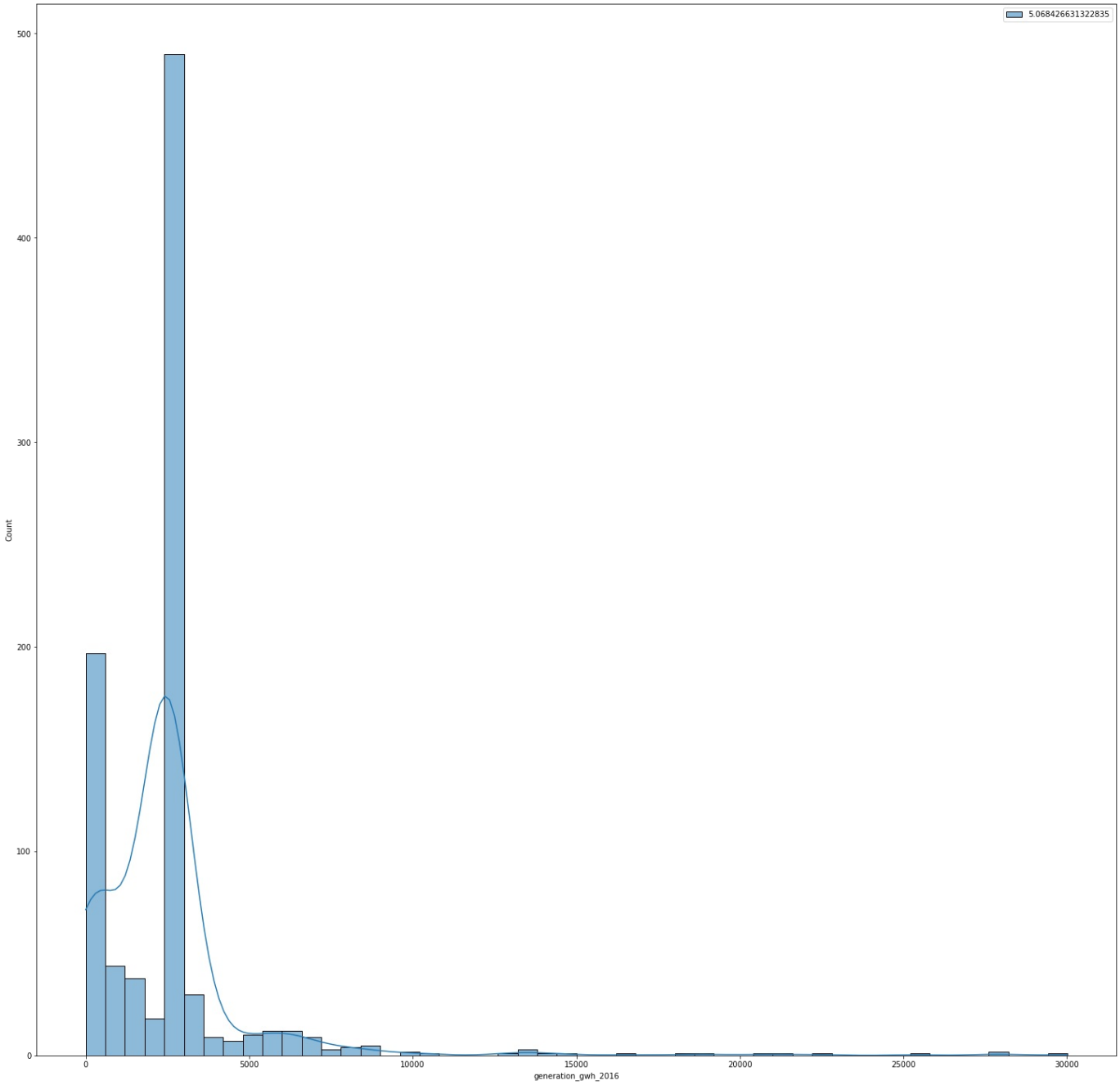
```
for i in a:
    sns.histplot(c[i], kde = True, bins = 50, label = c[i].skew())
    plt.legend(loc = 'upper right')
    plt.show()
```

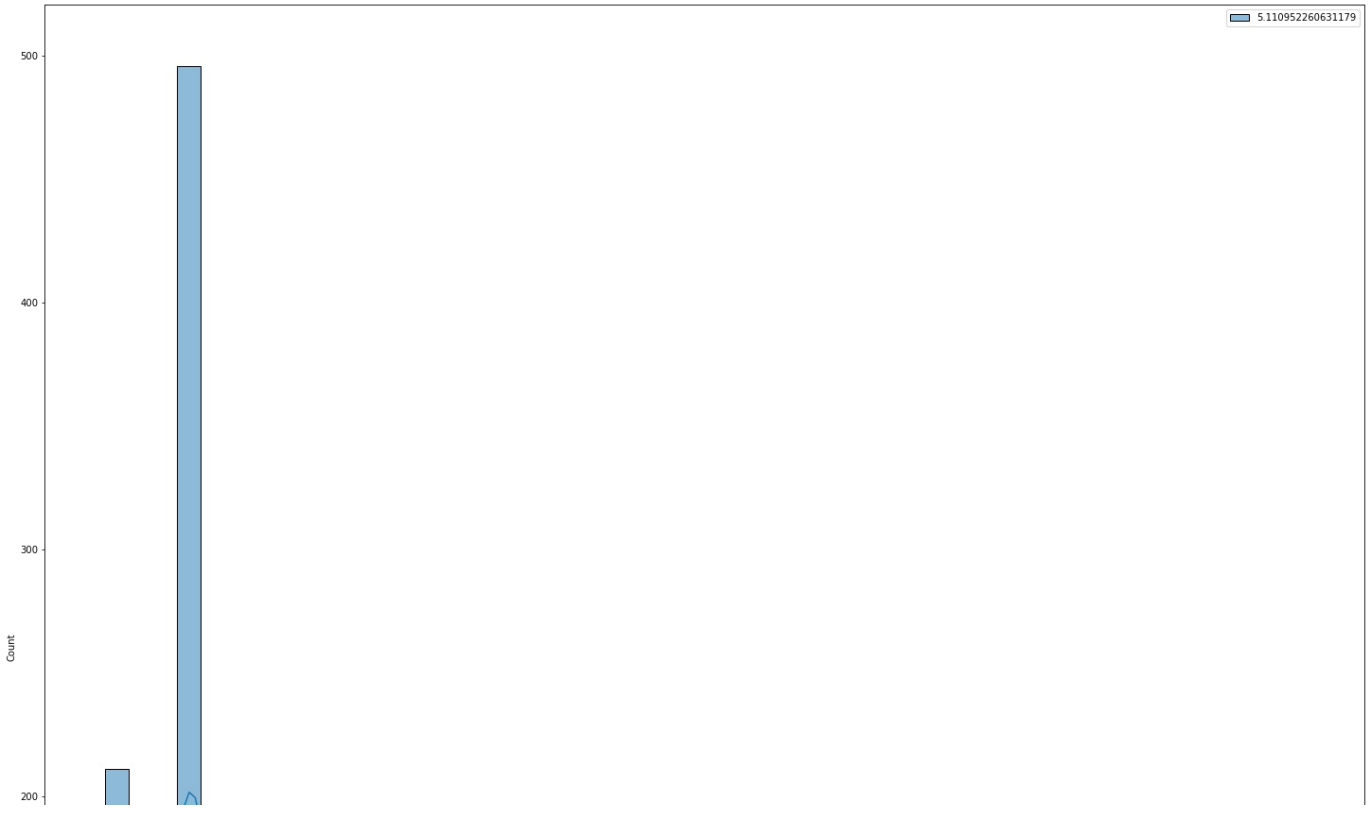


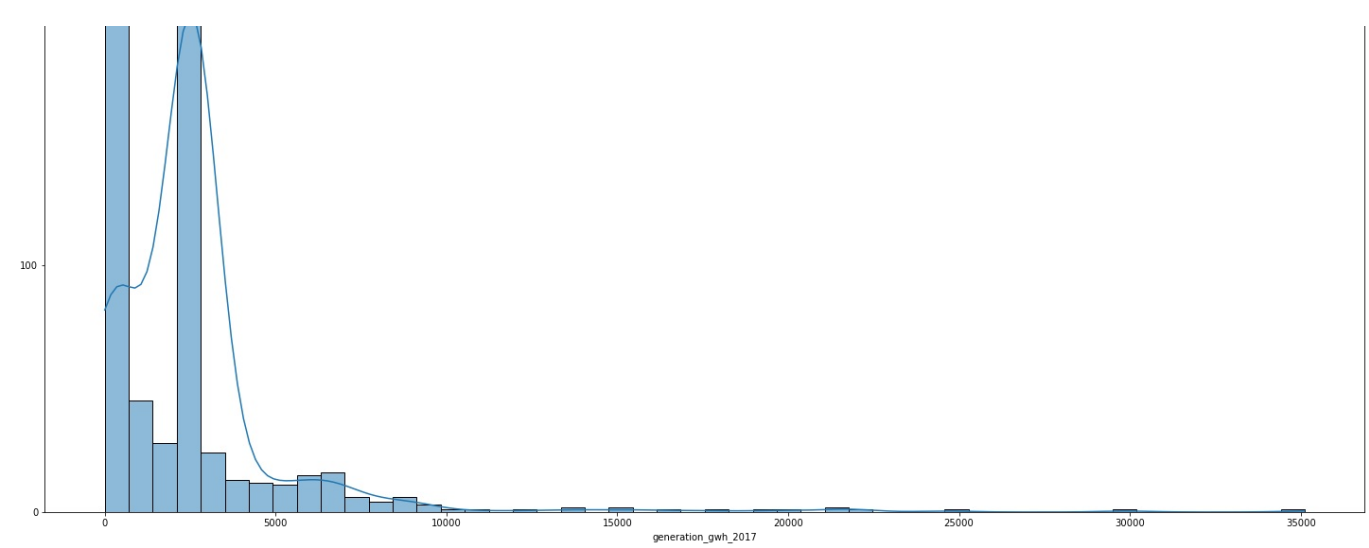


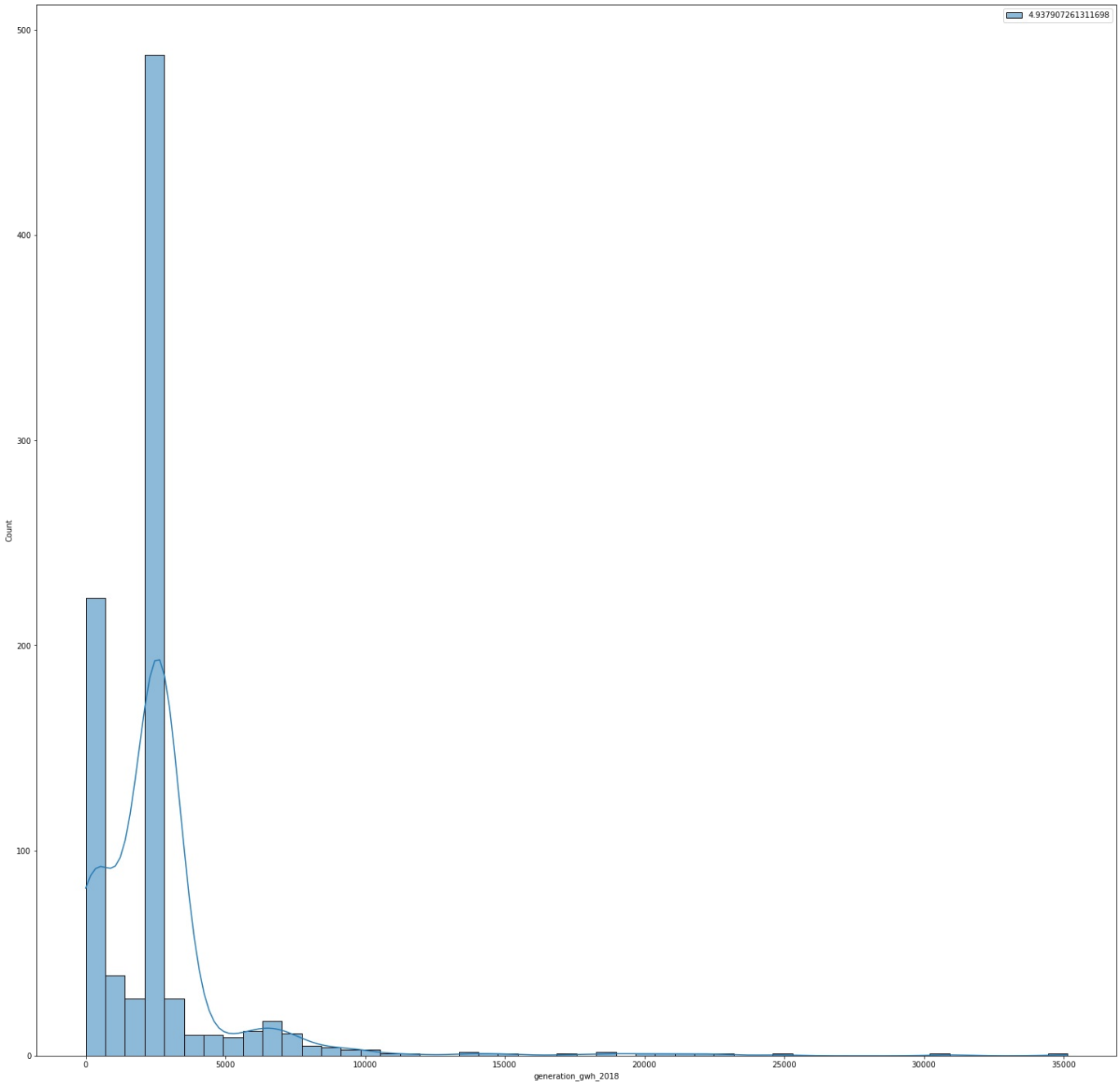

Count
300
200
100
0
0
5000
10000
15000
20000
25000
generation_gwh_2014


5.364134192254043
Count
500
400
300
200
100
0
0
5000
10000
15000
20000
25000
30000
generation_gwh_2015


5.068426631322835
Count
generation_gwh_2016


5.110952260631179
Count





In [356…

```
out_vars=['capacity_mw','generation_gwh_2014','generation_gwh_2015','generation_gwh_2016','generation_gwh_2017','
```

In [357…

```
def outlierTreat(x):
    upper = x.quantile(.75) + 1.5 * (x.quantile(.75) - x.quantile(.25))
    lower = x.quantile(.25) - 1.5 * (x.quantile(.75) - x.quantile(.25))
    return x.clip(lower, upper)
```

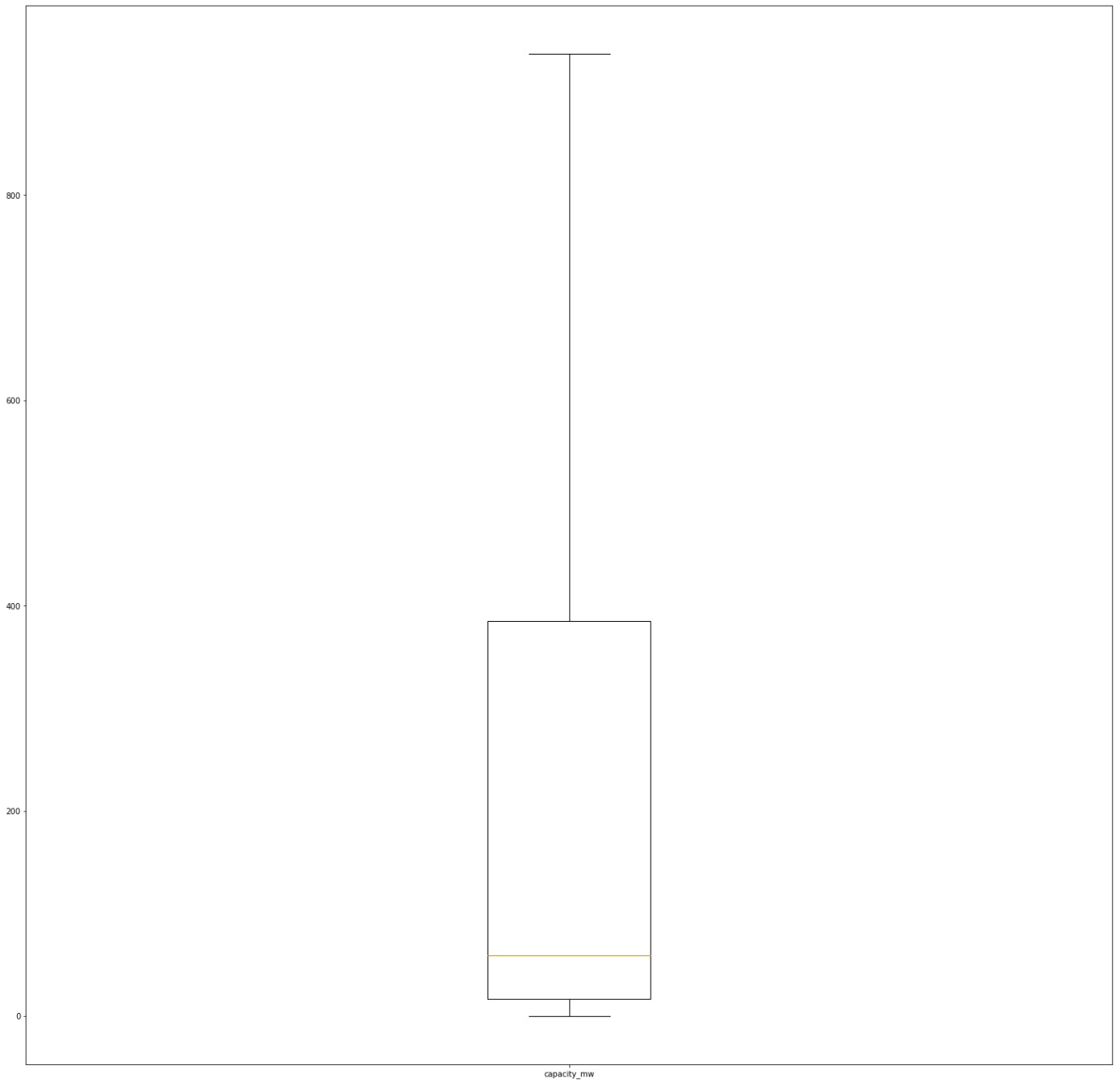
In [358…
```
c.loc[:, out_vars] = c.loc[:, out_vars].apply(outlierTreat)
c.loc[:, out_vars]
```

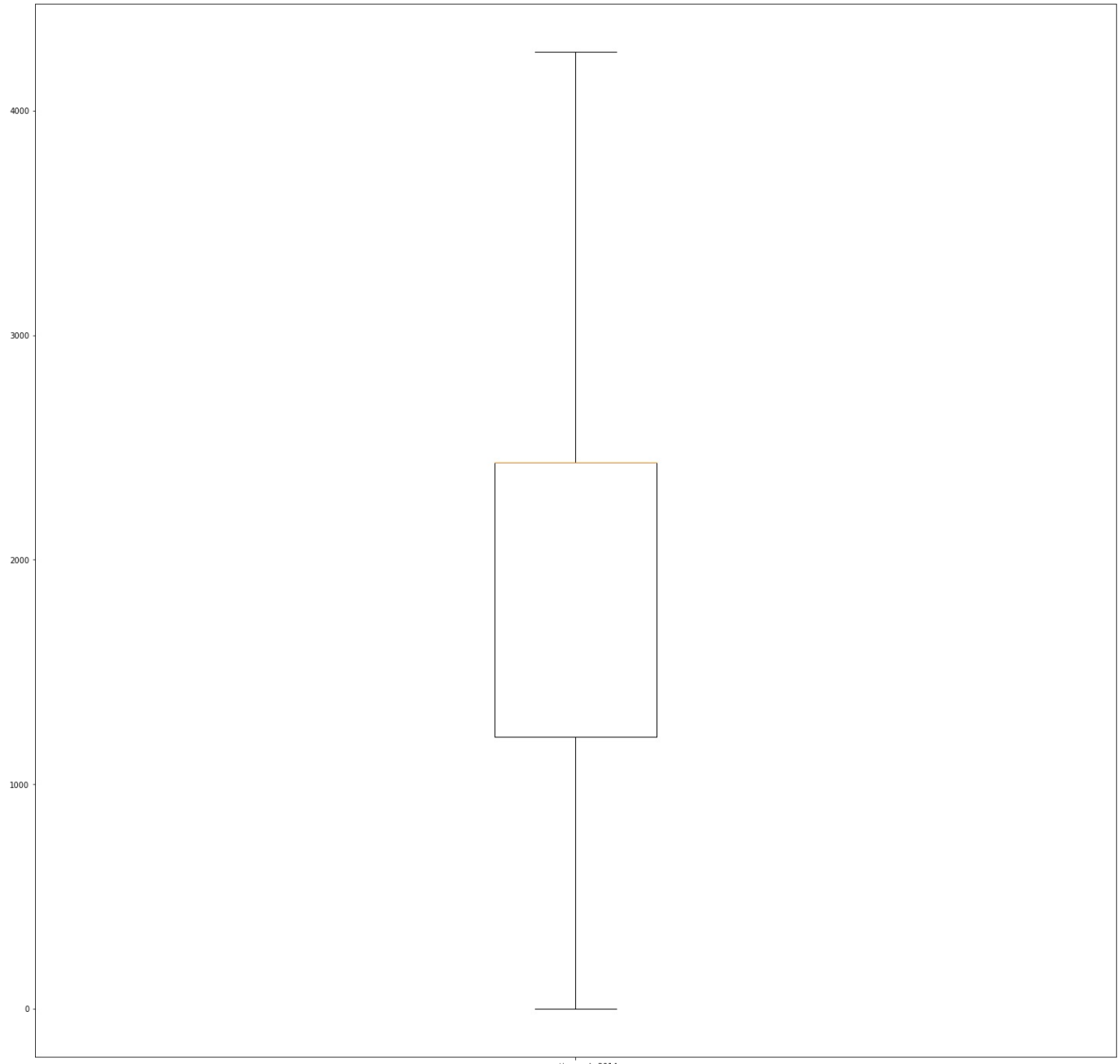
Out[358…

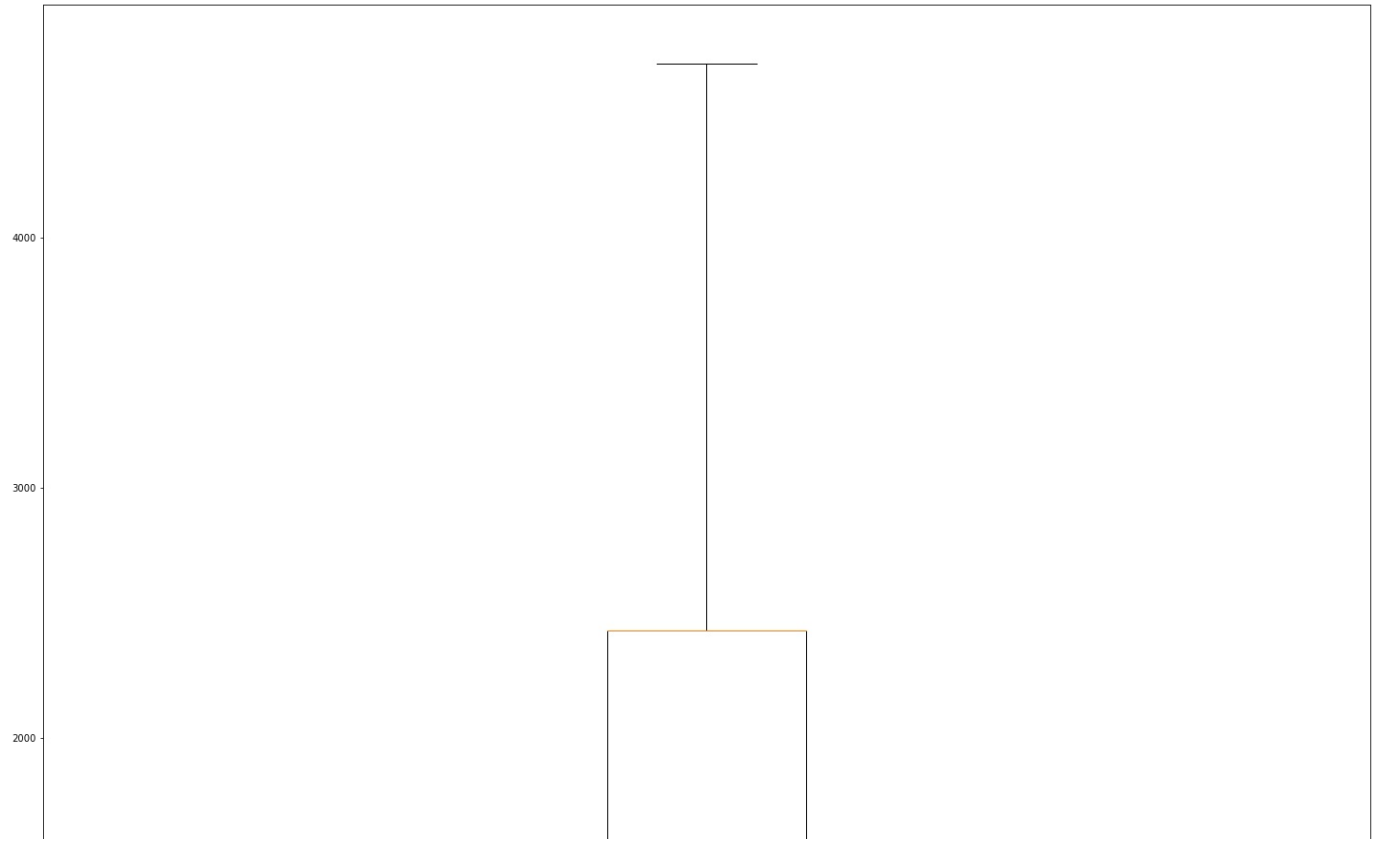
	capacity_mw	generation_gwh_2014	generation_gwh_2015	generation_gwh_2016	generation_gwh_2017	generation_gwh_2018
0	2.5000	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099
1	98.0000	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099
2	39.2000	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099
3	135.0000	617.789264	843.747000	886.004428	663.774500	626.239128
4	938.0375	3035.550000	4696.567365	4825.091826	5045.505986	5264.746736
...	...	...	...	...	...	...
902	938.0375	2431.823590	0.994875	233.596650	865.400000	686.500000
903	3.0000	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099
904	25.5000	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099
905	80.0000	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099
906	16.5000	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099

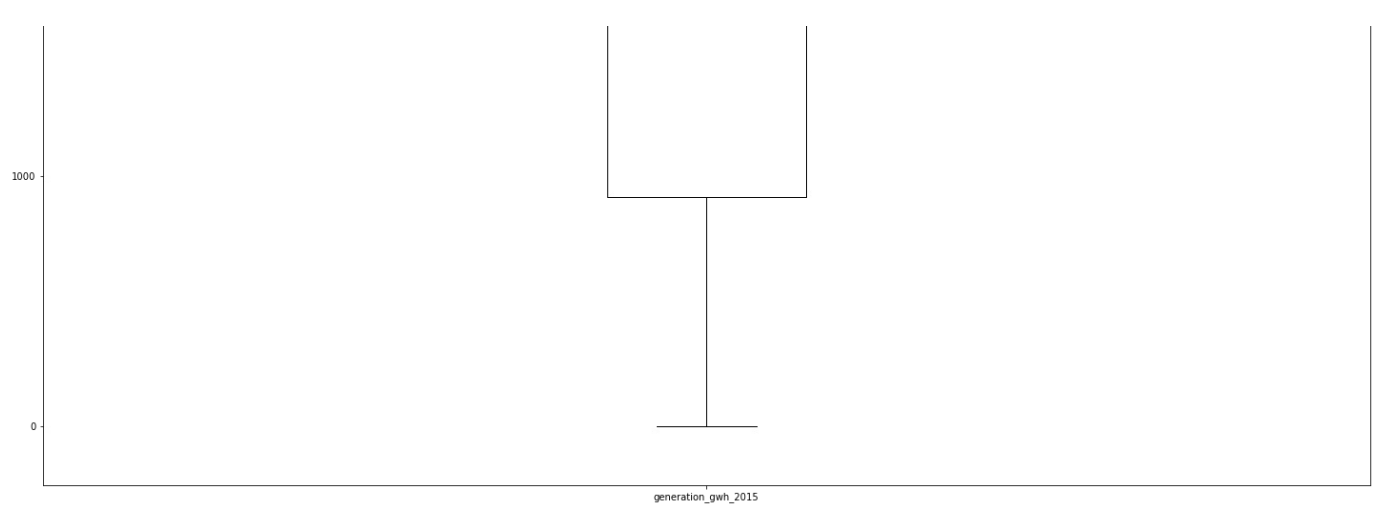
907 rows × 6 columns

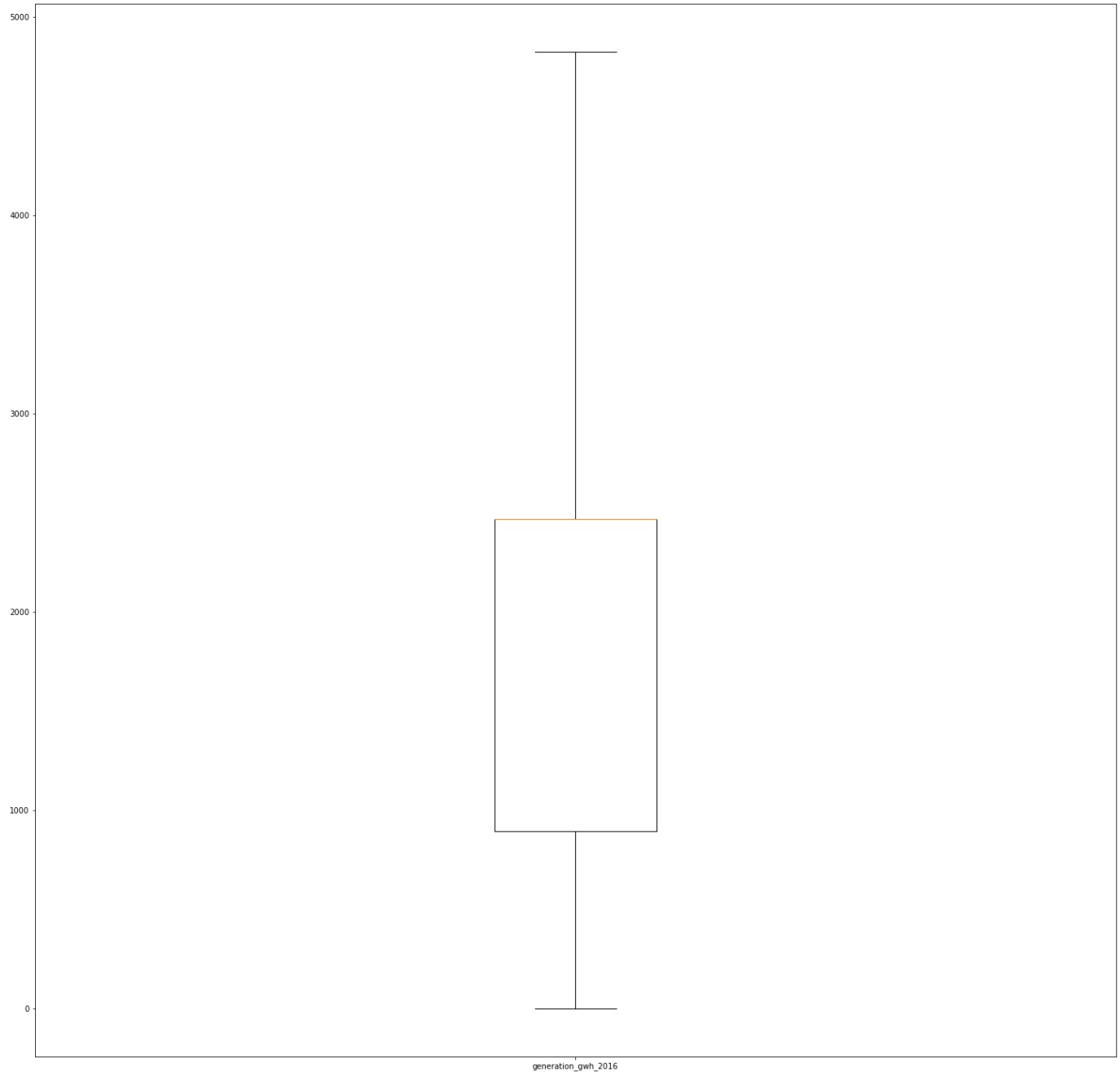
In [359…
```
# Using box plot for checking the presence of outliers.
for i in c:
    plt.boxplot(c[i], labels = [i])
    plt.show()
```

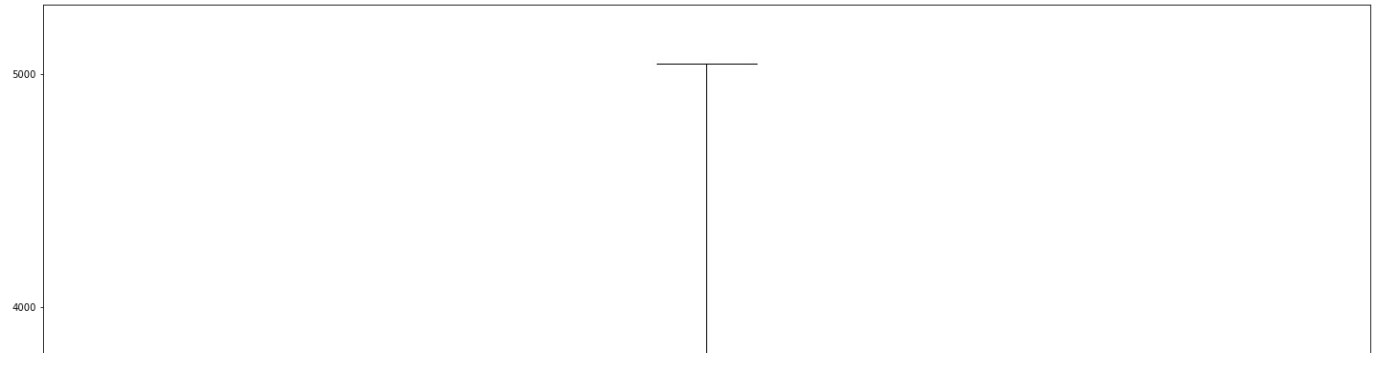


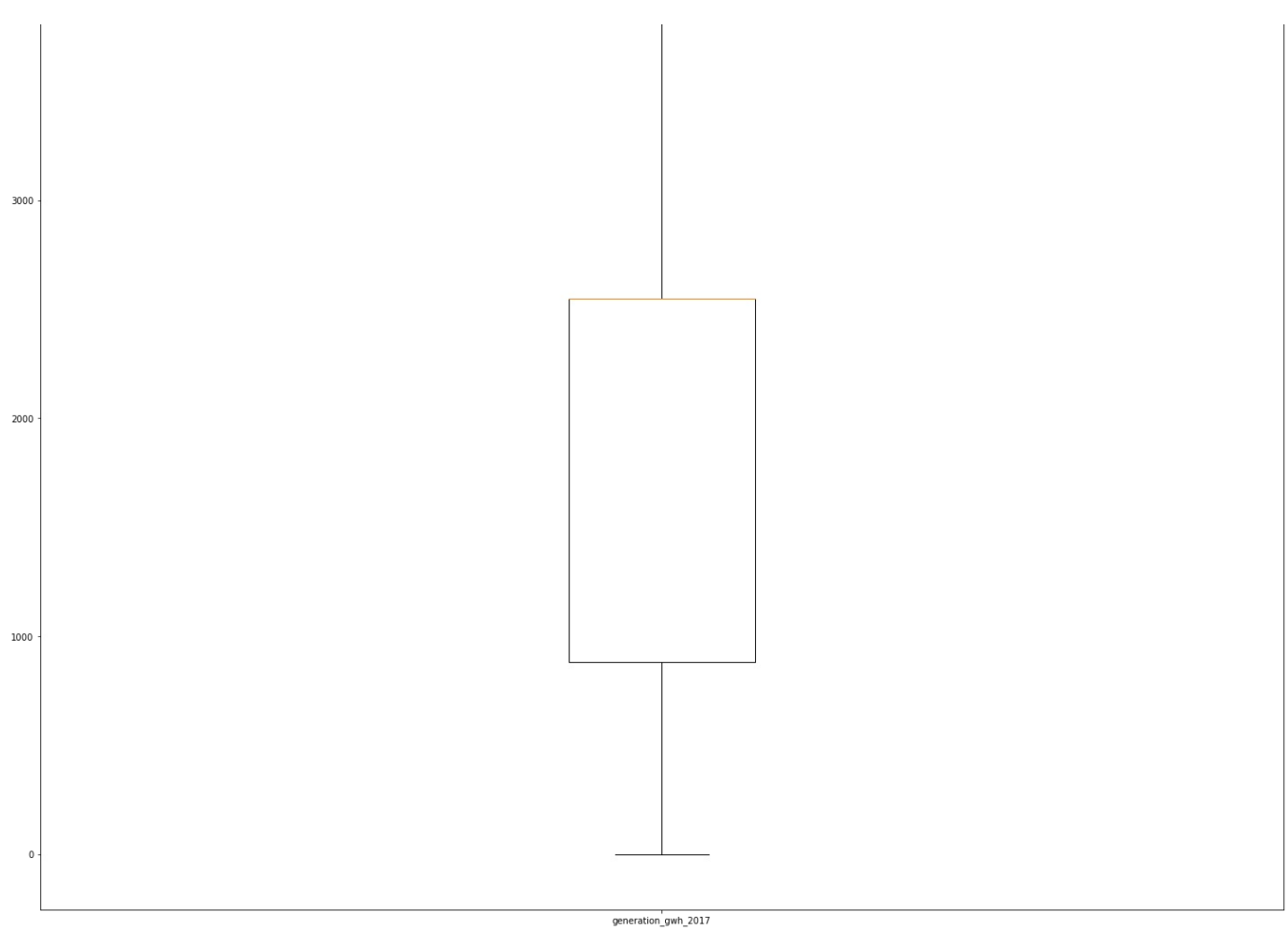

4000
3000
2000
1000
0
generation_gwh_2014

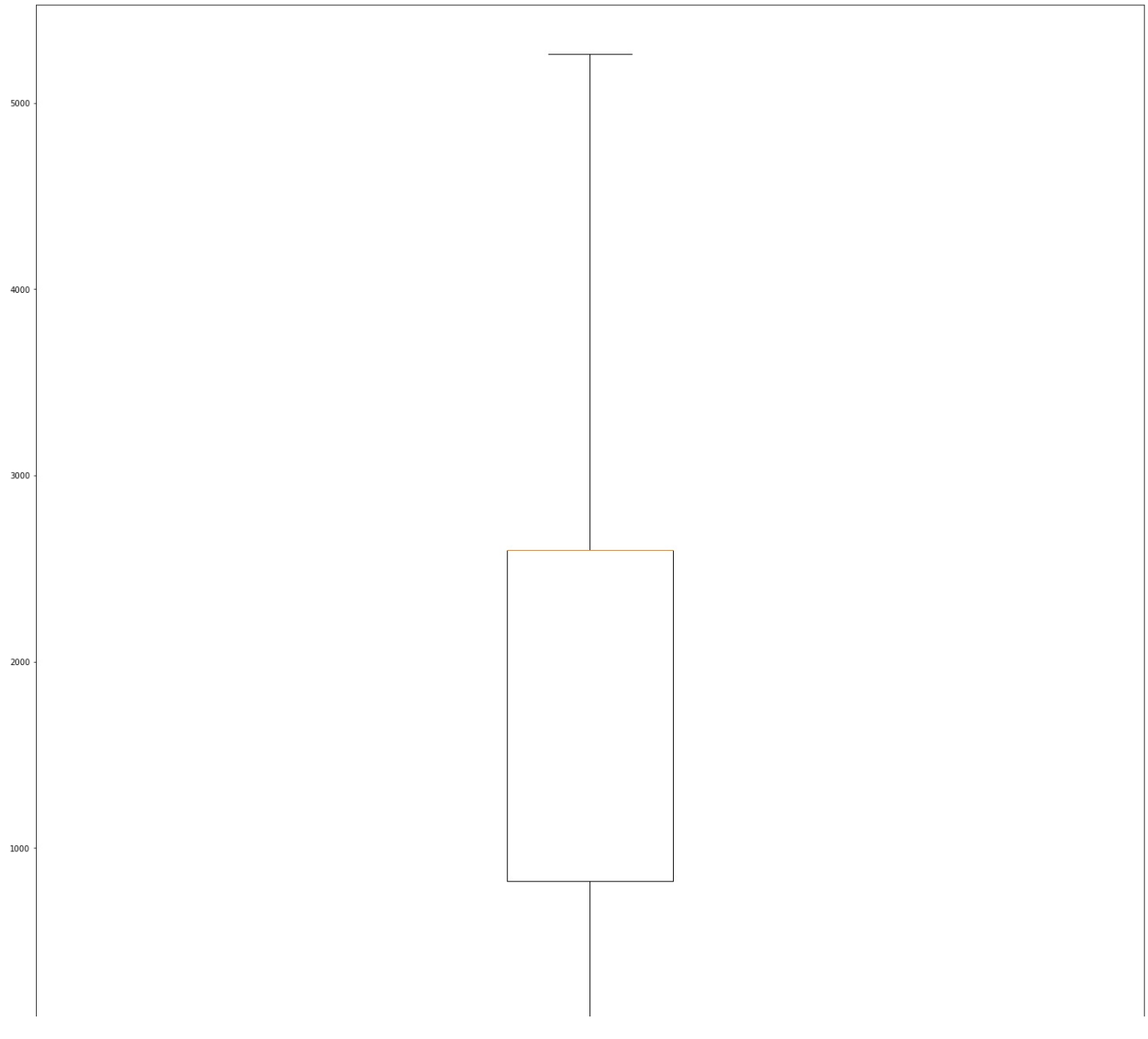

4000
3000
2000

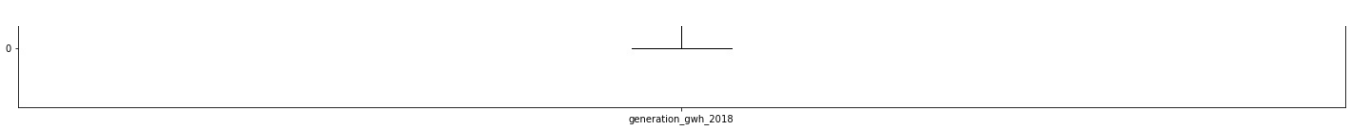

1000
0
generation_gwh_2015


5000
4000
3000
2000
1000
0
generation_gwh_2016


5000
4000


3000
2000
1000
0
generation_gwh_2017


5000
4000
3000
2000
1000



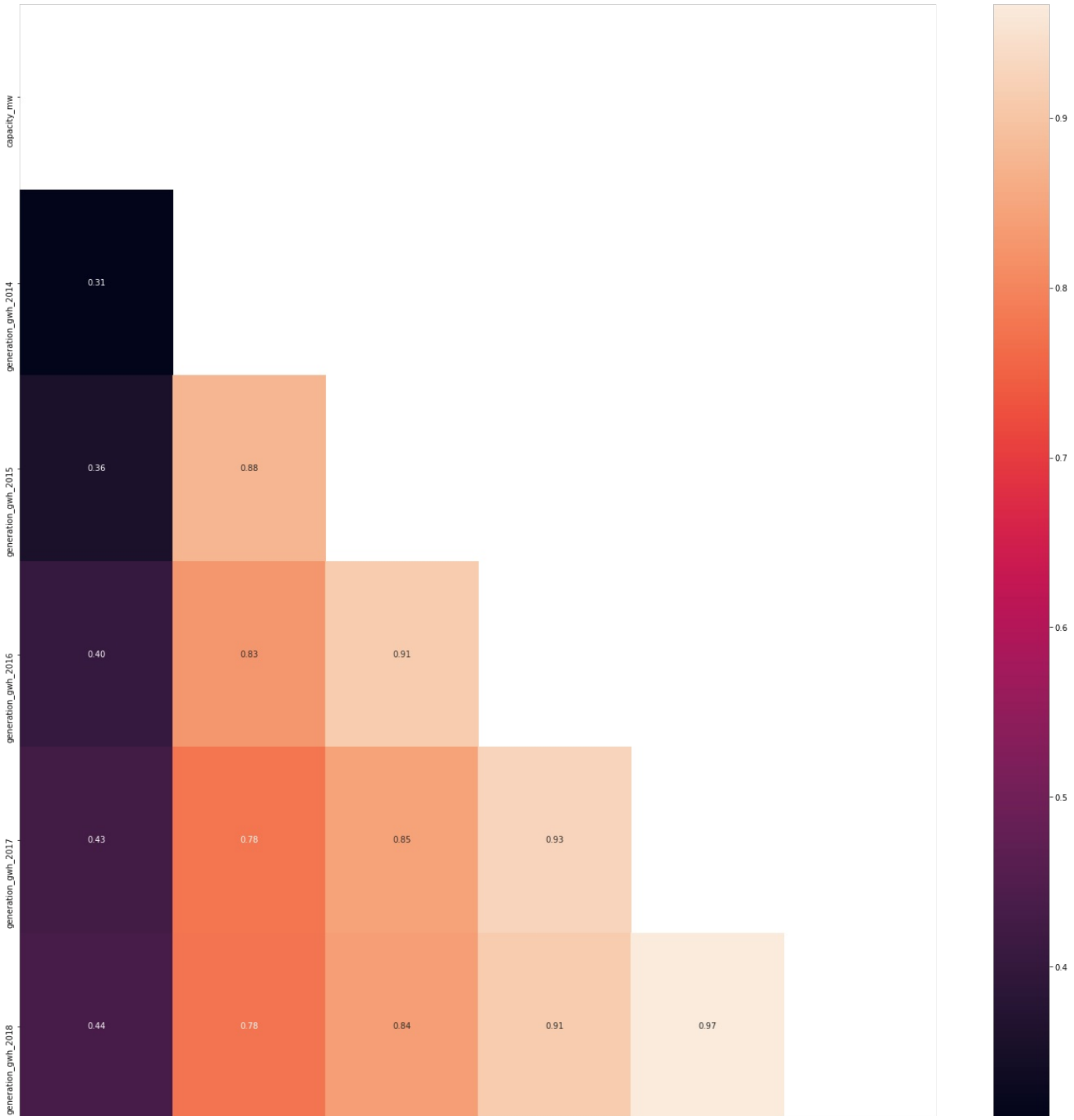
In [360…

```
## we have removed the outliers
```

In [361…

```
# Finding the correlation.
corr = c.corr()

# Setting the size of figure.
plt.rcParams['figure.figsize'] = (25, 25)

# Argument Trimming out the values above the main diagonal.
mask = np.triu(corr)

# Setting low correlation value to 0.
corr[(corr.values < 0.3) & (corr.values > -0.3)] = 0

# Plotting the heatmap.
sns.heatmap(corr, annot = True, fmt = '.2f', mask = mask)
```

Out[361… <AxesSubplot:>



In [362…
```
## exploring the categorical variables
```

In [363…
```
cat_vars = df.select_dtypes(include = ['object'])
cat_vars
```

Out[363…

	country	country_long	name	gppd_idnr	primary_fuel	other_fuel1	other_fuel2	owner	source	
0	IND	India	ACME Solar Tower	WRI1020239	Solar	NaN	NaN	Solar Paces	National Renewable Energy Laboratory	http://www.nrel.gov/csp/sol
1	IND	India	ADITYA CEMENT WORKS	WRI1019881	Coal	NaN	NaN	Ultratech Cement ltd	Ultratech Cement ltd	http://www.
2	IND	India	AES Saurashtra Windfarms	WRI1026669	Wind	NaN	NaN	AES	CDM	https://cdm.unfccc.int/Projects
3	IND	India	AGARTALA GT	IND0000001	Gas	NaN	NaN	NaN	Central Electricity Authority	
4	IND	India	AKALTARA TPP	IND0000002	Coal	Oil	NaN	NaN	Central Electricity Authority	
...	...	...	...	...	...	...	...	...	...	
902	IND	India	YERMARUS TPP	IND0000513	Coal	Oil	NaN	NaN	Central Electricity Authority	
903	IND	India	Yelesandra Solar Power Plant	WRI1026222	Solar	NaN	NaN	Karnataka Power Corporation Limited	Karnataka Power Corporation Limited	http:/
904	IND	India	Yelisirur wind power project	WRI1026776	Wind	NaN	NaN	NaN	CDM	https://cdm.unfccc.int/Projects
905	IND	India	ZAWAR MINES	WRI1019901	Coal	NaN	NaN	Hindustan Zinc ltd	Hindustan Zinc ltd	htt
906	IND	India	iEnergy Theni Wind Farm	WRI1026761	Wind	NaN	NaN	iEnergy Wind Farms	CDM	https://cdm.unfccc.int/Projects/

907 rows × 12 columns

In [364…
```
cat_vars.isnull().sum()
```

Out[364…
```
country                   0
country_long              0
name                      0
gppd_idnr                 0
primary_fuel              0
other_fuel1             709
other_fuel2             906
owner                   565
source                    0
url                       0
geolocation_source       19
generation_data_source  458
dtype: int64
```

In [365…
```
# Checking number of unique values in each columns
count = 1
for x in cat_vars:
    print(f'{count}. {x}: {cat_vars[x].nunique()}')
    print(f'{cat_vars[x].value_counts()}', end = '\n----------\n\n' )
    count += 1
```

```
1. country: 1
IND    907
Name: country, dtype: int64
----------
```

```
2. country_long: 1
India    907
Name: country_long, dtype: int64
----------

3. name: 907
Sadeipali - REHPL Solar Power Plant    1
SANGLI MIRAJ BIOMASS                   1
BARADARHA TPP                          1
MAHATMA SUGAR                          1
SOUTHERN REPL.                         1
                                      ..
CHANDA CEMENT WORKS                    1
MAIHAR CEMENT PLANT                    1
MOHAMAD PUR                            1
Kathauti 2 Solar Power Plant           1
TASHIDING                              1
Name: name, Length: 907, dtype: int64
----------

4. gppd_idnr: 907
IND0000381    1
IND0000331    1
IND0000186    1
WRI1026092    1
WRI1026761    1
             ..
WRI1026115    1
IND0000417    1
IND0000524    1
IND0000256    1
IND0000471    1
Name: gppd_idnr, Length: 907, dtype: int64
----------

5. primary_fuel: 8
Coal       258
Hydro      251
Solar      127
Wind       123
Gas         69
Biomass     50
Oil         20
Nuclear      9
Name: primary_fuel, dtype: int64
----------

6. other_fuel1: 3
Oil             195
Gas               2
Cogeneration      1
Name: other_fuel1, dtype: int64
----------

7. other_fuel2: 1
Oil    1
Name: other_fuel2, dtype: int64
----------

8. owner: 280
Jk Cement ltd                                    4
Acc Acc ltd                                      4
Sterling Agro Industries ltd.                    4
Powerica Limited                                 3
Shri Ssk ltd                                     3
                                                ..
Godavari Mills ltd                               1
Frost International Limited                      1
Solairedirect Projects India Private Limited     1
Saidham Overseas Private Limited                 1
West Coast Paper Mills Ltd.                      1
Name: owner, Length: 280, dtype: int64
----------

9. source: 191
Central Electricity Authority                519
CDM                                          124
Lancosola                                     10
National Renewable Energy Laboratory           8
National Thermal Power Corporation (NTPC)      6
                                            ...
Real Estate e                                  1
EMC Limited                                    1
```

```
Lokmangal Lokmangal group                       1
Maral Overseas ltd                              1
West Coast Paper Mills Ltd.                     1
Name: source, Length: 191, dtype: int64
----------

10. url: 304
http://www.cea.nic.in/
519
http://www.lancosolar.com/pdfs/rajasthan-pv-project-details.pdf
7
http://www.ntpc.co.in
6
http://viainfotech.biz/Biomass/theme5/document/green_market/REC-project-list.pdf
5
http://energy.rajasthan.gov.in/content/dam/raj/energy/common/Details%20of%20commissioned%20Solar%20Projects%20.pd
f                               4

...
https://cdm.unfccc.int/filestorage/w/m/64TXH0Y1V9ZCISBKO3F758PEQNUJDR.pdf/PDD__V-2_5_19_10_2012.pdf?t=akh8b2pkZTF
tfDAP0pu4sjZSao0P-GV-Qzqn       1
http://documents.worldbank.org/curated/en/442061468041961880/pdf/multi-page.pdf
1
http://www.tradeindia.com/Seller-6835496-Shri-Dudhganga-Vedganga-SSK-Ltd-/
1
http://www.thoratsugar.com/
1
https://cdm.unfccc.int/Projects/DB/LRQA%20Ltd1346322352.66/view
1
Name: url, Length: 304, dtype: int64
----------

11. geolocation_source: 3
WRI                                   765
Industry About                        119
National Renewable Energy Laboratory    4
Name: geolocation_source, dtype: int64
----------

12. generation_data_source: 1
Central Electricity Authority    449
Name: generation_data_source, dtype: int64
----------
```

In [366…

```
##dropping the unnecessary columns
```

In [367…

```
d=cat_vars.drop(['country','other_fuel2','url','gppd_idnr','owner','name'],axis=1)
d
```

Out[367…

	country_long	primary_fuel	other_fuel1	source	geolocation_source	generation_data_source
0	India	Solar	NaN	National Renewable Energy Laboratory	National Renewable Energy Laboratory	NaN
1	India	Coal	NaN	Ultratech Cement ltd	WRI	NaN
2	India	Wind	NaN	CDM	WRI	NaN
3	India	Gas	NaN	Central Electricity Authority	WRI	Central Electricity Authority
4	India	Coal	Oil	Central Electricity Authority	WRI	Central Electricity Authority
...	...	...	...	...	...	...
902	India	Coal	Oil	Central Electricity Authority	WRI	Central Electricity Authority
903	India	Solar	NaN	Karnataka Power Corporation Limited	Industry About	NaN
904	India	Wind	NaN	CDM	WRI	NaN
905	India	Coal	NaN	Hindustan Zinc ltd	WRI	NaN
906	India	Wind	NaN	CDM	WRI	NaN

907 rows × 6 columns

In [368…

```
d
```

Out[368…

	country_long	primary_fuel	other_fuel1	source	geolocation_source	generation_data_source
0	India	Solar	NaN	National Renewable Energy Laboratory	National Renewable Energy Laboratory	NaN
1	India	Coal	NaN	Ultratech Cement ltd	WRI	NaN
2	India	Wind	NaN	CDM	WRI	NaN
3	India	Gas	NaN	Central Electricity Authority	WRI	Central Electricity Authority
4	India	Coal	Oil	Central Electricity Authority	WRI	Central Electricity Authority
...	...	...	...	...	...	...
902	India	Coal	Oil	Central Electricity Authority	WRI	Central Electricity Authority
903	India	Solar	NaN	Karnataka Power Corporation Limited	Industry About	NaN
904	India	Wind	NaN	CDM	WRI	NaN
905	India	Coal	NaN	Hindustan Zinc ltd	WRI	NaN
906	India	Wind	NaN	CDM	WRI	NaN

907 rows × 6 columns

In [369…

```
d.isnull().sum()
```

Out[369…

```
country_long                0
primary_fuel                0
other_fuel1               709
source                      0
geolocation_source         19
generation_data_source    458
dtype: int64
```

In [370…

```
## filling the null values
```

In [371…

```
d_=d.fillna(d.mode().iloc[0])
d_
```

Out[371…

	country_long	primary_fuel	other_fuel1	source	geolocation_source	generation_data_source
0	India	Solar	Oil	National Renewable Energy Laboratory	National Renewable Energy Laboratory	Central Electricity Authority
1	India	Coal	Oil	Ultratech Cement ltd	WRI	Central Electricity Authority
2	India	Wind	Oil	CDM	WRI	Central Electricity Authority
3	India	Gas	Oil	Central Electricity Authority	WRI	Central Electricity Authority
4	India	Coal	Oil	Central Electricity Authority	WRI	Central Electricity Authority
...	...	...	...	...	...	...
902	India	Coal	Oil	Central Electricity Authority	WRI	Central Electricity Authority
903	India	Solar	Oil	Karnataka Power Corporation Limited	Industry About	Central Electricity Authority
904	India	Wind	Oil	CDM	WRI	Central Electricity Authority
905	India	Coal	Oil	Hindustan Zinc ltd	WRI	Central Electricity Authority
906	India	Wind	Oil	CDM	WRI	Central Electricity Authority

907 rows × 6 columns

In [372…

```
d_.isnull().sum()
```

Out[372…

```
country_long              0
primary_fuel              0
```

```
other_fuel1               0
source                    0
geolocation_source        0
generation_data_source    0
dtype: int64
```

In [373…
```
## null values has been removed
```

In [374…
```
cat_data = d_.copy()
cat_data = pd.get_dummies(d_, drop_first = True) ## numerical features to continuos features
cat_data
```

Out[374…

	primary_fuel_Coal	primary_fuel_Gas	primary_fuel_Hydro	primary_fuel_Nuclear	primary_fuel_Oil	primary_fuel_Solar	primary_fuel_Wind	oth
0	0	0	0	0	0	1	0	
1	1	0	0	0	0	0	0	
2	0	0	0	0	0	0	1	
3	0	1	0	0	0	0	0	
4	1	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	
902	1	0	0	0	0	0	0	
903	0	0	0	0	0	1	0	
904	0	0	0	0	0	0	1	
905	1	0	0	0	0	0	0	
906	0	0	0	0	0	0	1	

907 rows × 201 columns

In [375…
```
# Combining Numerical and Categorical data.
final_data = pd.concat([c, cat_data], axis = 1)
final_data
```

Out[375…

	capacity_mw	generation_gwh_2014	generation_gwh_2015	generation_gwh_2016	generation_gwh_2017	generation_gwh_2018	primary_fuel_(
0	2.5000	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099	
1	98.0000	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099	
2	39.2000	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099	
3	135.0000	617.789264	843.747000	886.004428	663.774500	626.239128	
4	938.0375	3035.550000	4696.567365	4825.091826	5045.505986	5264.746736	
...	...	...	...	...	...	...	
902	938.0375	2431.823590	0.994875	233.596650	865.400000	686.500000	
903	3.0000	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099	
904	25.5000	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099	
905	80.0000	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099	
906	16.5000	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099	

907 rows × 207 columns

In [376…
```
x=final_data.drop(['capacity_mw'],axis=1)
x
```

Out[376…

	generation_gwh_2014	generation_gwh_2015	generation_gwh_2016	generation_gwh_2017	generation_gwh_2018	primary_fuel_Coal	primary_
0	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099	0	
1	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099	1	
2	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099	0	
3	617.789264	843.747000	886.004428	663.774500	626.239128	0	
4	3035.550000	4696.567365	4825.091826	5045.505986	5264.746736	1	
...	...	...	...	...	...	...	
902	2431.823590	0.994875	233.596650	865.400000	686.500000	1	
903	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099	0	
904	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099	0	
905	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099	1	
906	2431.823590	2428.226946	2467.936859	2547.759305	2600.804099	0	

907 rows × 206 columns

In [377…
```
y=final_data['capacity_mw']
y
```

Out[377…
```
0        2.5000
1       98.0000
2       39.2000
3      135.0000
4      938.0375
         ...
902    938.0375
903      3.0000
904     25.5000
905     80.0000
906     16.5000
Name: capacity_mw, Length: 907, dtype: float64
```

In [378…
```
from sklearn.preprocessing import StandardScaler
```

In [379…
```
st=StandardScaler()
```

In [380…
```
st.fit_transform(x)
```

Out[380…
```
array([[ 0.33156797,  0.29135674,  0.28386702, ..., -0.03322277,
        15.0249792 , -2.52467498],
       [ 0.33156797,  0.29135674,  0.28386702, ..., -0.03322277,
        -0.06655583,  0.39609059],
       [ 0.33156797,  0.29135674,  0.28386702, ..., -0.03322277,
        -0.06655583,  0.39609059],
       ...,
       [ 0.33156797,  0.29135674,  0.28386702, ..., -0.03322277,
        -0.06655583,  0.39609059],
       [ 0.33156797,  0.29135674,  0.28386702, ..., -0.03322277,
        -0.06655583,  0.39609059],
       [ 0.33156797,  0.29135674,  0.28386702, ..., -0.03322277,
        -0.06655583,  0.39609059]])
```

In [381…
```
from sklearn.model_selection import train_test_split,cross_val_score
```

In [382…
```
#importing models
from sklearn.neighbors import KNeighborsRegressor
from sklearn.linear_model import LinearRegression,Lasso,Ridge,ElasticNet
from sklearn.svm import SVR
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor,AdaBoostRegressor,GradientBoostingRegressor
```

In [383…
```
from sklearn.metrics import r2_score,mean_absolute_error,mean_squared_error
```

In [384…
```
x_train,x_test,y_train,y_test=train_test_split(x,y,random_state=41,test_size=0.25)
```

In [385…
```
kn=KNeighborsRegressor()
```

In [386…
```
kn.fit(x_train,y_train)
```

Out[386… KNeighborsRegressor()

In [387…
```
y_pred=kn.predict(x_test)
```

In [388…
```
mean_absolute_error(y_test,y_pred)
```

Out[388… 85.43816431718064

In [389…
```
mean_squared_error(y_test,y_pred)
```

Out[389… 24992.956824032994

In [390…
```
r2_score(y_test,y_pred)
```

Out[390… 0.798297748625687

In [391…
```
## linear regression
```

In [392…
```
lr=LinearRegression()
```

In [393…
```
lr.fit(x_train,y_train)
```

Out[393… LinearRegression()

In [394…
```
y_pred=lr.predict(x_test)
```

In [395…
```
r2_score(y_test,y_pred)
```

Out[395… 0.6943821910402179

In [396…
```
mean_squared_error(y_test,y_pred)
```

Out[396… 37869.14946136364

In [397…
```
##SVR
```

In [398…
```
sv=SVR()
```

In [399…
```
sv.fit(x_train,y_train)
```

Out[399… SVR()

In [400…
```
y_pred=sv.predict(x_test)
```

In [401…
```
r2_score(y_test,y_pred)
```

Out[401… -0.3347746589790095

In [402…
```
## Decision tree regressor
```

In [403…
```
dt=DecisionTreeRegressor()
```

In [404…
```
dt.fit(x_train,y_train)
```

Out[404… DecisionTreeRegressor()

In [405…
```
y_pred=dt.predict(x_test)
```

In [406…
```
r2_score(y_test,y_pred)
```

Out[406… 0.7948003788601744

In [407…
```
mean_squared_error(y_test,y_pred)
```

Out[407… 25426.31644670237

In [408…
```
mean_absolute_error(y_test,y_pred)
```

Out[408… 73.16966308028417

In [409…
```
## Randomforest
```

In [410…
```
rf=RandomForestRegressor()
```

In [411…
```
rf.fit(x_train,y_train)
```

Out[411… RandomForestRegressor()

In [412…
```
y_pred=rf.predict(x_test)
```

In [413…
```
mean_absolute_error(y_test,y_pred)
```

Out[413… 65.61095911953018

In [414…
```
r2_score(y_test,y_pred)
```

Out[414… 0.84009982731287

In [415…
```
mean_squared_error(y_test,y_pred)
```

Out[415… 19813.254859057073

In [416…
```
## gradientBoosting regressor
```

In [417…
```
gb=GradientBoostingRegressor()
```

In [418…
```
gb.fit(x_train,y_train)
```

Out[418… GradientBoostingRegressor()

In [419…
```
y_pred=gb.predict(x_test)
```

In [420…
```
mean_absolute_error(y_test,y_pred)
```

Out[420… 75.55876223595511

In [421…
```
r2_score(y_test,y_pred)
```

Out[421… 0.8189936983678565

In [422…
```
mean_squared_error(y_test,y_pred)
```

Out[422… 22428.518525431668

In [423…
```
## ada boost regressor
```

In [424…
```
ab=AdaBoostRegressor()
```

In [425…
```
ab.fit(x_train,y_train)
```

Out[425… AdaBoostRegressor()

In [426…
```
y_pred=ab.predict(x_test)
```

In [427…
```
mean_absolute_error(y_test,y_pred)
```

Out[427… 102.44868518941618

In [428…
```
r2_score(y_test,y_pred)
```

Out[428… 0.783238584338263

In [429…
```
mean_squared_error(y_test,y_pred)
```

Out[429… 26858.94017462607

In [430…
```
## looks like only two i.e. randomforest and gradientboosting works good
```

In [431…
```
##hyper parameter tuning
```

In [432…
```
from sklearn.model_selection import RandomizedSearchCV
```

In [433…
```
params={'n_estimators':[100,200,300,400,500,600,700],'min_samples_split':[1,2,3,4],'min_samples_leaf':[1,2,3,4],'
```

In [434…
```
g=RandomizedSearchCV(RandomForestRegressor(),params,cv=10)
```

In [435…
```
g.fit(x_train,y_train)
```

Out[435…
```
RandomizedSearchCV(cv=10, estimator=RandomForestRegressor(),
                   param_distributions={'max_depth': [None, 1, 2, 3, 4, 5, 6, 7,
                                                      8],
                                        'min_samples_leaf': [1, 2, 3, 4],
                                        'min_samples_split': [1, 2, 3, 4],
                                        'n_estimators': [100, 200, 300, 400,
                                                         500, 600, 700]})
```

In [436…
```
z=g.best_params_
z
```

Out[436…
```
{'n_estimators': 300,
 'min_samples_split': 4,
 'min_samples_leaf': 2,
 'max_depth': None}
```

In [437…
```
m=RandomForestRegressor(n_estimators= 400,
 min_samples_split= 3,
 min_samples_leaf=2,
 max_depth=None)
```

In [438…
```
m.fit(x_train,y_train)
```

Out[438…
```
RandomForestRegressor(min_samples_leaf=2, min_samples_split=3, n_estimators=400)
```

In [439…
```
y_test=m.predict(x_test)
```

In [440…
```
mean_absolute_error(y_test,y_pred)
```

Out[440…
```
60.152284504987286
```

In [441…
```
r2_score(y_test,y_pred)*100
```

Out[441…
```
92.40999823581879
```

In [442…
```
mean_squared_error(y_test,y_pred)
```

Out[442…
```
8205.37423000566
```

In [443…
```
## random forest regressor working efficiently
```

In [444…
```
##conclusion
```

In [445…
```
import numpy as np
```

In [446…
```
a=np.array(y_test)
```

In [447…
```
predicted=np.array(m.predict(x_test))
```

In [448…
```
df_com=pd.DataFrame({'actual':a,'pred':predicted},index=range(len(a)))
```

In [449…
```
df_com
```

Out[449…

	actual	pred
0	112.630114	112.630114
1	659.647165	659.647165
2	668.731248	668.731248
3	46.538538	46.538538
4	212.480868	212.480868
...	...	...
222	80.599960	80.599960
223	337.387669	337.387669
224	33.324610	33.324610
225	33.324610	33.324610
226	125.229955	125.229955

227 rows × 2 columns

In [450…
```
## now we will see the classifier
```

In [451…
```
df.head()
```

Out[451…

	country	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	other_fuel2	...	year_of_capacity
0	IND	India	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	NaN	NaN	...	
1	IND	India	ADITYA CEMENT WORKS	WRI1019881	98.0	24.7663	74.6090	Coal	NaN	NaN	...	
2	IND	India	AES Saurashtra Windfarms	WRI1026669	39.2	21.9038	69.3732	Wind	NaN	NaN	...	
3	IND	India	AGARTALA GT	IND0000001	135.0	23.8712	91.3602	Gas	NaN	NaN	...	2
4	IND	India	AKALTARA TPP	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	NaN	...	2

5 rows × 27 columns

In [452…
```
df.isnull().sum()
```

Out[452…
```
country                   0
country_long              0
name                      0
gppd_idnr                 0
capacity_mw               0
latitude                 46
longitude                46
primary_fuel              0
other_fuel1             709
other_fuel2             906
other_fuel3             907
commissioning_year      380
owner                   565
source                    0
url                       0
geolocation_source       19
```

```
wepp_id                    907
year_of_capacity_data      388
generation_gwh_2013        907
generation_gwh_2014        509
generation_gwh_2015        485
generation_gwh_2016        473
generation_gwh_2017        467
generation_gwh_2018        459
generation_gwh_2019        907
generation_data_source     458
estimated_generation_gwh   907
dtype: int64
```

In [453…
```
L=df.drop(['country','other_fuel2','other_fuel3','wepp_id','generation_gwh_2013','generation_gwh_2019','estimated
```

In [454…
```
L
```

Out[454…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	commissioning_year	owner	s
0	India	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	NaN	2011.0	Solar Paces	Na Rene E Labo
1	India	ADITYA CEMENT WORKS	WRI1019881	98.0	24.7663	74.6090	Coal	NaN	NaN	Ultratech Cement ltd	Ult Cem
2	India	AES Saurashtra Windfarms	WRI1026669	39.2	21.9038	69.3732	Wind	NaN	NaN	AES	
3	India	AGARTALA GT	IND0000001	135.0	23.8712	91.3602	Gas	NaN	2004.0	NaN	C Ele Au
4	India	AKALTARA TPP	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	2015.0	NaN	C Ele Au
...	...	...	...	...	...	...	...	...	...	...	
902	India	YERMARUS TPP	IND0000513	1600.0	16.2949	77.3568	Coal	Oil	2016.0	NaN	C Ele Au
903	India	Yelesandra Solar Power Plant	WRI1026222	3.0	12.8932	78.1654	Solar	NaN	NaN	Karnataka Power Corporation Limited	Kar Corp L
904	India	Yelisirur wind power project	WRI1026776	25.5	15.2758	75.5811	Wind	NaN	NaN	NaN	
905	India	ZAWAR MINES	WRI1019901	80.0	24.3500	73.7477	Coal	NaN	NaN	Hindustan Zinc ltd	Hin Z
906	India	iEnergy Theni Wind Farm	WRI1026761	16.5	9.9344	77.4768	Wind	NaN	NaN	iEnergy Wind Farms	

907 rows × 20 columns

In [455…
```
## filling the null values
```

In [456…
```
L['generation_gwh_2017'].fillna(L['generation_gwh_2017'].mean(),inplace=True)
```

In [457…
```
L
```

Out[457…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	commissioning_year	owner	s
0	India	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	NaN	2011.0	Solar Paces	Na Rene E Labo
1	India	ADITYA CEMENT WORKS	WRI1019881	98.0	24.7663	74.6090	Coal	NaN	NaN	Ultratech Cement ltd	Ult Cem
2	India	AES Saurashtra Windfarms	WRI1026669	39.2	21.9038	69.3732	Wind	NaN	NaN	AES	

3	India	AGARTALA GT	IND0000001	135.0	23.8712	91.3602	Gas	NaN	2004.0	NaN	G Ele Au
4	India	AKALTARA TPP	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	2015.0	NaN	G Ele Au
...	...	...	...	...	...	...	...	...	...	...	
902	India	YERMARUS TPP	IND0000513	1600.0	16.2949	77.3568	Coal	Oil	2016.0	NaN	G Ele Au
903	India	Yelesandra Solar Power Plant	WRI1026222	3.0	12.8932	78.1654	Solar	NaN	NaN	Karnataka Power Corporation Limited	Karn Corpo L
904	India	Yelisirur wind power project	WRI1026776	25.5	15.2758	75.5811	Wind	NaN	NaN	NaN	
905	India	ZAWAR MINES	WRI1019901	80.0	24.3500	73.7477	Coal	NaN	NaN	Hindustan Zinc ltd	Hind Z
906	India	iEnergy Theni Wind Farm	WRI1026761	16.5	9.9344	77.4768	Wind	NaN	NaN	iEnergy Wind Farms	

907 rows × 20 columns

In [458…
```
L['generation_gwh_2018'].fillna(L['generation_gwh_2018'].mean(),inplace=True)
```

In [459…
```
L
```

Out[459…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	commissioning_year	owner	s
0	India	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	NaN	2011.0	Solar Paces	Na Rene E Labo
1	India	ADITYA CEMENT WORKS	WRI1019881	98.0	24.7663	74.6090	Coal	NaN	NaN	Ultratech Cement ltd	Ult Cem
2	India	AES Saurashtra Windfarms	WRI1026669	39.2	21.9038	69.3732	Wind	NaN	NaN	AES	
3	India	AGARTALA GT	IND0000001	135.0	23.8712	91.3602	Gas	NaN	2004.0	NaN	G Ele Au
4	India	AKALTARA TPP	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	2015.0	NaN	G Ele Au
...	...	...	...	...	...	...	...	...	...	...	
902	India	YERMARUS TPP	IND0000513	1600.0	16.2949	77.3568	Coal	Oil	2016.0	NaN	G Ele Au
903	India	Yelesandra Solar Power Plant	WRI1026222	3.0	12.8932	78.1654	Solar	NaN	NaN	Karnataka Power Corporation Limited	Karn Corpo L
904	India	Yelisirur wind power project	WRI1026776	25.5	15.2758	75.5811	Wind	NaN	NaN	NaN	
905	India	ZAWAR MINES	WRI1019901	80.0	24.3500	73.7477	Coal	NaN	NaN	Hindustan Zinc ltd	Hind Z
906	India	iEnergy Theni Wind Farm	WRI1026761	16.5	9.9344	77.4768	Wind	NaN	NaN	iEnergy Wind Farms	

907 rows × 20 columns

In [460…
```
L['generation_gwh_2015'].fillna(L['generation_gwh_2015'].mean(),inplace=True)
```

In [461…
```
L
```

Out[461…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	commissioning_year	owner	s

0	India	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	NaN	2011.0	Solar Paces	Na Rene E Labo
1	India	ADITYA CEMENT WORKS	WRI1019881	98.0	24.7663	74.6090	Coal	NaN	NaN	Ultratech Cement ltd	Ult Cem
2	India	AES Saurashtra Windfarms	WRI1026669	39.2	21.9038	69.3732	Wind	NaN	NaN	AES	
3	India	AGARTALA GT	IND0000001	135.0	23.8712	91.3602	Gas	NaN	2004.0	NaN	C Ele Au
4	India	AKALTARA TPP	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	2015.0	NaN	C Ele Au
...	...	...	...	...	...	...	...	...	...	...	
902	India	YERMARUS TPP	IND0000513	1600.0	16.2949	77.3568	Coal	Oil	2016.0	NaN	C Ele Au
903	India	Yelesandra Solar Power Plant	WRI1026222	3.0	12.8932	78.1654	Solar	NaN	NaN	Karnataka Power Corporation Limited	Karı Corpo L
904	India	Yelisirur wind power project	WRI1026776	25.5	15.2758	75.5811	Wind	NaN	NaN	NaN	
905	India	ZAWAR MINES	WRI1019901	80.0	24.3500	73.7477	Coal	NaN	NaN	Hindustan Zinc ltd	Hinc Z
906	India	iEnergy Theni Wind Farm	WRI1026761	16.5	9.9344	77.4768	Wind	NaN	NaN	iEnergy Wind Farms	

907 rows × 20 columns

In [462…
```
L['generation_gwh_2014'].fillna(L['generation_gwh_2014'].mean(),inplace=True)
```

In [463…
```
L
```

Out[463…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	commissioning_year	owner	s
0	India	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	NaN	2011.0	Solar Paces	Na Rene E Labo
1	India	ADITYA CEMENT WORKS	WRI1019881	98.0	24.7663	74.6090	Coal	NaN	NaN	Ultratech Cement ltd	Ult Cem
2	India	AES Saurashtra Windfarms	WRI1026669	39.2	21.9038	69.3732	Wind	NaN	NaN	AES	
3	India	AGARTALA GT	IND0000001	135.0	23.8712	91.3602	Gas	NaN	2004.0	NaN	C Ele Au
4	India	AKALTARA TPP	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	2015.0	NaN	C Ele Au
...	...	...	...	...	...	...	...	...	...	...	
902	India	YERMARUS TPP	IND0000513	1600.0	16.2949	77.3568	Coal	Oil	2016.0	NaN	C Ele Au
903	India	Yelesandra Solar Power Plant	WRI1026222	3.0	12.8932	78.1654	Solar	NaN	NaN	Karnataka Power Corporation Limited	Karı Corpo L
904	India	Yelisirur wind power project	WRI1026776	25.5	15.2758	75.5811	Wind	NaN	NaN	NaN	
905	India	ZAWAR MINES	WRI1019901	80.0	24.3500	73.7477	Coal	NaN	NaN	Hindustan Zinc ltd	Hinc Z
906	India	iEnergy Theni Wind Farm	WRI1026761	16.5	9.9344	77.4768	Wind	NaN	NaN	iEnergy Wind Farms	

907 rows × 20 columns

In [464…
```
L['generation_gwh_2016'].fillna(L['generation_gwh_2016'].mean(),inplace=True)
```

In [465…
```
L
```

Out[465…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	commissioning_year	owner	s
0	India	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	NaN	2011.0	Solar Paces	Na Rene E Labo
1	India	ADITYA CEMENT WORKS	WRI1019881	98.0	24.7663	74.6090	Coal	NaN	NaN	Ultratech Cement ltd	Ult Cem
2	India	AES Saurashtra Windfarms	WRI1026669	39.2	21.9038	69.3732	Wind	NaN	NaN	AES	
3	India	AGARTALA GT	IND0000001	135.0	23.8712	91.3602	Gas	NaN	2004.0	NaN	C Ele Au
4	India	AKALTARA TPP	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	2015.0	NaN	C Ele Au
...	...	...	...	...	...	...	...	...	...	...	
902	India	YERMARUS TPP	IND0000513	1600.0	16.2949	77.3568	Coal	Oil	2016.0	NaN	C Ele Au
903	India	Yelesandra Solar Power Plant	WRI1026222	3.0	12.8932	78.1654	Solar	NaN	NaN	Karnataka Power Corporation Limited	Kar Corp L
904	India	Yelisirur wind power project	WRI1026776	25.5	15.2758	75.5811	Wind	NaN	NaN	NaN	
905	India	ZAWAR MINES	WRI1019901	80.0	24.3500	73.7477	Coal	NaN	NaN	Hindustan Zinc ltd	Hin Z
906	India	iEnergy Theni Wind Farm	WRI1026761	16.5	9.9344	77.4768	Wind	NaN	NaN	iEnergy Wind Farms	

907 rows × 20 columns

In [466…
```
L['latitude'].fillna(L['latitude'].mean(),inplace=True)
```

In [467…
```
L
```

Out[467…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	commissioning_year	owner	s
0	India	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	NaN	2011.0	Solar Paces	Na Rene E Labo
1	India	ADITYA CEMENT WORKS	WRI1019881	98.0	24.7663	74.6090	Coal	NaN	NaN	Ultratech Cement ltd	Ult Cem
2	India	AES Saurashtra Windfarms	WRI1026669	39.2	21.9038	69.3732	Wind	NaN	NaN	AES	
3	India	AGARTALA GT	IND0000001	135.0	23.8712	91.3602	Gas	NaN	2004.0	NaN	C Ele Au
4	India	AKALTARA TPP	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	2015.0	NaN	C Ele Au
...	...	...	...	...	...	...	...	...	...	...	
902	India	YERMARUS TPP	IND0000513	1600.0	16.2949	77.3568	Coal	Oil	2016.0	NaN	C Ele Au
903	India	Yelesandra Solar Power Plant	WRI1026222	3.0	12.8932	78.1654	Solar	NaN	NaN	Karnataka Power Corporation Limited	Kar Corp L
904	India	Yelisirur wind power	WRI1026776	25.5	15.2758	75.5811	Wind	NaN	NaN	NaN	

		project									
905	India	ZAWAR MINES	WRI1019901	80.0	24.3500	73.7477	Coal	NaN	NaN	Hindustan Zinc ltd	Hinc Z
906	India	iEnergy Theni Wind Farm	WRI1026761	16.5	9.9344	77.4768	Wind	NaN	NaN	iEnergy Wind Farms	

907 rows × 20 columns

In [468…
```
L['longitude'].fillna(L['longitude'].mean(),inplace=True)
```

In [469…
```
L
```

Out[469…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	commissioning_year	owner	s
0	India	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	NaN	2011.0	Solar Paces	Na Rene E Labo
1	India	ADITYA CEMENT WORKS	WRI1019881	98.0	24.7663	74.6090	Coal	NaN	NaN	Ultratech Cement ltd	Ult Cem
2	India	AES Saurashtra Windfarms	WRI1026669	39.2	21.9038	69.3732	Wind	NaN	NaN	AES	
3	India	AGARTALA GT	IND0000001	135.0	23.8712	91.3602	Gas	NaN	2004.0	NaN	C Ele Au
4	India	AKALTARA TPP	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	2015.0	NaN	C Ele Au
...	...	...	...	...	...	...	...	...	...	...	
902	India	YERMARUS TPP	IND0000513	1600.0	16.2949	77.3568	Coal	Oil	2016.0	NaN	C Ele Au
903	India	Yelesandra Solar Power Plant	WRI1026222	3.0	12.8932	78.1654	Solar	NaN	NaN	Karnataka Power Corporation Limited	Kar Corp L
904	India	Yelisirur wind power project	WRI1026776	25.5	15.2758	75.5811	Wind	NaN	NaN	NaN	
905	India	ZAWAR MINES	WRI1019901	80.0	24.3500	73.7477	Coal	NaN	NaN	Hindustan Zinc ltd	Hinc Z
906	India	iEnergy Theni Wind Farm	WRI1026761	16.5	9.9344	77.4768	Wind	NaN	NaN	iEnergy Wind Farms	

907 rows × 20 columns

In [470…
```
L['commissioning_year'].fillna(L['commissioning_year'].mean(),inplace=True)
```

In [471…
```
L
```

Out[471…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	commissioning_year	owner	s
0	India	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	NaN	2011.000000	Solar Paces	Na Rene E Labo
1	India	ADITYA CEMENT WORKS	WRI1019881	98.0	24.7663	74.6090	Coal	NaN	1997.091082	Ultratech Cement ltd	Ult Cem
2	India	AES Saurashtra Windfarms	WRI1026669	39.2	21.9038	69.3732	Wind	NaN	1997.091082	AES	
3	India	AGARTALA GT	IND0000001	135.0	23.8712	91.3602	Gas	NaN	2004.000000	NaN	C Ele Au
4	India	AKALTARA TPP	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	2015.000000	NaN	C Ele Au

...	...	...	...	...	...	...	...	...	...	...
902	India	YERMARUS TPP	IND0000513	1600.0	16.2949	77.3568	Coal	Oil	2016.000000	NaN
903	India	Yelesandra Solar Power Plant	WRI1026222	3.0	12.8932	78.1654	Solar	NaN	1997.091082	Karnataka Power Corporation Limited
904	India	Yelisirur wind power project	WRI1026776	25.5	15.2758	75.5811	Wind	NaN	1997.091082	NaN
905	India	ZAWAR MINES	WRI1019901	80.0	24.3500	73.7477	Coal	NaN	1997.091082	Hindustan Zinc ltd
906	India	iEnergy Theni Wind Farm	WRI1026761	16.5	9.9344	77.4768	Wind	NaN	1997.091082	iEnergy Wind Farms

907 rows × 20 columns

In [472…
```
L['year_of_capacity_data'].fillna(L['year_of_capacity_data'].mean(),inplace=True)
```

In [473…
```
L
```

Out[473…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	commissioning_year	owner
0	India	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	NaN	2011.000000	Solar Paces
1	India	ADITYA CEMENT WORKS	WRI1019881	98.0	24.7663	74.6090	Coal	NaN	1997.091082	Ultratech Cement ltd
2	India	AES Saurashtra Windfarms	WRI1026669	39.2	21.9038	69.3732	Wind	NaN	1997.091082	AES
3	India	AGARTALA GT	IND0000001	135.0	23.8712	91.3602	Gas	NaN	2004.000000	NaN
4	India	AKALTARA TPP	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	2015.000000	NaN
...	...	...	...	...	...	...	...	...	...	...
902	India	YERMARUS TPP	IND0000513	1600.0	16.2949	77.3568	Coal	Oil	2016.000000	NaN
903	India	Yelesandra Solar Power Plant	WRI1026222	3.0	12.8932	78.1654	Solar	NaN	1997.091082	Karnataka Power Corporation Limited
904	India	Yelisirur wind power project	WRI1026776	25.5	15.2758	75.5811	Wind	NaN	1997.091082	NaN
905	India	ZAWAR MINES	WRI1019901	80.0	24.3500	73.7477	Coal	NaN	1997.091082	Hindustan Zinc ltd
906	India	iEnergy Theni Wind Farm	WRI1026761	16.5	9.9344	77.4768	Wind	NaN	1997.091082	iEnergy Wind Farms

907 rows × 20 columns

In [474…
```
L.isnull().sum()
```

Out[474…
```
country_long              0
name                      0
gppd_idnr                 0
capacity_mw               0
latitude                  0
longitude                 0
primary_fuel              0
other_fuel1             709
commissioning_year        0
owner                   565
source                    0
```

```
url                          0
geolocation_source          19
year_of_capacity_data        0
generation_gwh_2014          0
generation_gwh_2015          0
generation_gwh_2016          0
generation_gwh_2017          0
generation_gwh_2018          0
generation_data_source     458
dtype: int64
```

In [475…

```
L=L.fillna(L.mode().iloc[0])
L
```

Out[475…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	commissioning_year	owner	s
0	India	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	Oil	2011.000000	Solar Paces	Na Rene E Labo
1	India	ADITYA CEMENT WORKS	WRI1019881	98.0	24.7663	74.6090	Coal	Oil	1997.091082	Ultratech Cement ltd	Ult Cem
2	India	AES Saurashtra Windfarms	WRI1026669	39.2	21.9038	69.3732	Wind	Oil	1997.091082	AES	
3	India	AGARTALA GT	IND0000001	135.0	23.8712	91.3602	Gas	Oil	2004.000000	Acc Acc ltd	C Ele Au
4	India	AKALTARA TPP	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	2015.000000	Acc Acc ltd	C Ele Au
...	...	...	...	...	...	...	...	...	...	...	
902	India	YERMARUS TPP	IND0000513	1600.0	16.2949	77.3568	Coal	Oil	2016.000000	Acc Acc ltd	C Ele Au
903	India	Yelesandra Solar Power Plant	WRI1026222	3.0	12.8932	78.1654	Solar	Oil	1997.091082	Karnataka Power Corporation Limited	Kar Corp L
904	India	Yelisirur wind power project	WRI1026776	25.5	15.2758	75.5811	Wind	Oil	1997.091082	Acc Acc ltd	
905	India	ZAWAR MINES	WRI1019901	80.0	24.3500	73.7477	Coal	Oil	1997.091082	Hindustan Zinc ltd	Hin Z
906	India	iEnergy Theni Wind Farm	WRI1026761	16.5	9.9344	77.4768	Wind	Oil	1997.091082	iEnergy Wind Farms	

907 rows × 20 columns

In [476…

```
L.isnull().sum()
```

Out[476…

```
country_long             0
name                     0
gppd_idnr                0
capacity_mw              0
latitude                 0
longitude                0
primary_fuel             0
other_fuel1              0
commissioning_year       0
owner                    0
source                   0
url                      0
geolocation_source       0
year_of_capacity_data    0
generation_gwh_2014      0
generation_gwh_2015      0
generation_gwh_2016      0
generation_gwh_2017      0
generation_gwh_2018      0
generation_data_source   0
dtype: int64
```

In [477…
```
## we have removed all the null values
```

In [478…
```
L=L.drop(['owner','url','commissioning_year'],axis=1)
L
```

Out[478…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	source	geolocation_source	year_c
0	India	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	Oil	National Renewable Energy Laboratory	National Renewable Energy Laboratory	
1	India	ADITYA CEMENT WORKS	WRI1019881	98.0	24.7663	74.6090	Coal	Oil	Ultratech Cement ltd	WRI	
2	India	AES Saurashtra Windfarms	WRI1026669	39.2	21.9038	69.3732	Wind	Oil	CDM	WRI	
3	India	AGARTALA GT	IND0000001	135.0	23.8712	91.3602	Gas	Oil	Central Electricity Authority	WRI	
4	India	AKALTARA TPP	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	Central Electricity Authority	WRI	
...	...	...	...	...	...	...	...	...	...	...	
902	India	YERMARUS TPP	IND0000513	1600.0	16.2949	77.3568	Coal	Oil	Central Electricity Authority	WRI	
903	India	Yelesandra Solar Power Plant	WRI1026222	3.0	12.8932	78.1654	Solar	Oil	Karnataka Power Corporation Limited	Industry About	
904	India	Yelisirur wind power project	WRI1026776	25.5	15.2758	75.5811	Wind	Oil	CDM	WRI	
905	India	ZAWAR MINES	WRI1019901	80.0	24.3500	73.7477	Coal	Oil	Hindustan Zinc ltd	WRI	
906	India	iEnergy Theni Wind Farm	WRI1026761	16.5	9.9344	77.4768	Wind	Oil	CDM	WRI	

907 rows × 17 columns

In [479…
```
L
```

Out[479…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	source	geolocation_source	year_c
0	India	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	Oil	National Renewable Energy Laboratory	National Renewable Energy Laboratory	
1	India	ADITYA CEMENT WORKS	WRI1019881	98.0	24.7663	74.6090	Coal	Oil	Ultratech Cement ltd	WRI	
2	India	AES Saurashtra Windfarms	WRI1026669	39.2	21.9038	69.3732	Wind	Oil	CDM	WRI	
3	India	AGARTALA GT	IND0000001	135.0	23.8712	91.3602	Gas	Oil	Central Electricity Authority	WRI	
4	India	AKALTARA TPP	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	Central Electricity Authority	WRI	
...	...	...	...	...	...	...	...	...	...	...	
902	India	YERMARUS TPP	IND0000513	1600.0	16.2949	77.3568	Coal	Oil	Central Electricity Authority	WRI	
903	India	Yelesandra Solar Power Plant	WRI1026222	3.0	12.8932	78.1654	Solar	Oil	Karnataka Power Corporation Limited	Industry About	
904	India	Yelisirur wind power project	WRI1026776	25.5	15.2758	75.5811	Wind	Oil	CDM	WRI	
905	India	ZAWAR MINES	WRI1019901	80.0	24.3500	73.7477	Coal	Oil	Hindustan Zinc ltd	WRI	

906	India	iEnergy Theni Wind Farm	WRI1026761	16.5	9.9344	77.4768	Wind	Oil	CDM	WRI

907 rows × 17 columns

In [480…

```
# Checking number of unique values in each columns
count = 1
for x in L :
    print(f'{count}. {x}: {L[x].nunique()}')
    print(f'{L[x].value_counts()}', end = '\n----------\n\n' )
    count += 1
```

```
1. country_long: 1
India    907
Name: country_long, dtype: int64
----------

2. name: 907
Sadeipali - REHPL Solar Power Plant    1
SANGLI MIRAJ BIOMASS                   1
BARADARHA TPP                          1
MAHATMA SUGAR                          1
SOUTHERN REPL.                         1
                                      ..
CHANDA CEMENT WORKS                    1
MAIHAR CEMENT PLANT                    1
MOHAMAD PUR                            1
Kathauti 2 Solar Power Plant           1
TASHIDING                              1
Name: name, Length: 907, dtype: int64
----------

3. gppd_idnr: 907
IND0000381    1
IND0000331    1
IND0000186    1
WRI1026092    1
WRI1026761    1
             ..
WRI1026115    1
IND0000417    1
IND0000524    1
IND0000256    1
IND0000471    1
Name: gppd_idnr, Length: 907, dtype: int64
----------

4. capacity_mw: 361
5.0       39
10.0      22
15.0      20
600.0     20
1200.0    19
          ..
192.0      1
27.3       1
26.4       1
68.8       1
19.7       1
Name: capacity_mw, Length: 361, dtype: int64
----------

5. latitude: 837
21.197918    46
24.191700     3
19.000400     3
15.261500     2
13.245000     2
             ..
9.087000      1
20.909900     1
17.238700     1
23.559400     1
16.597300     1
Name: latitude, Length: 837, dtype: int64
----------

6. longitude: 828
77.464907    46
```

```
71.691700     4
72.898300     3
71.691800     3
75.898800     3
             ..
91.811400     1
80.126400     1
76.113700     1
74.644700     1
79.574800     1
Name: longitude, Length: 828, dtype: int64
----------

7. primary_fuel: 8
Coal       258
Hydro      251
Solar      127
Wind       123
Gas         69
Biomass     50
Oil         20
Nuclear      9
Name: primary_fuel, dtype: int64
----------

8. other_fuel1: 3
Oil             904
Gas               2
Cogeneration      1
Name: other_fuel1, dtype: int64
----------

9. source: 191
Central Electricity Authority                519
CDM                                          124
Lancosola                                     10
National Renewable Energy Laboratory           8
National Thermal Power Corporation (NTPC)      6
                                             ...
Real Estate e                                  1
EMC Limited                                    1
Lokmangal Lokmangal group                      1
Maral Overseas ltd                             1
West Coast Paper Mills Ltd.                    1
Name: source, Length: 191, dtype: int64
----------

10. geolocation_source: 3
WRI                                    784
Industry About                         119
National Renewable Energy Laboratory     4
Name: geolocation_source, dtype: int64
----------

11. year_of_capacity_data: 1
2019.0    907
Name: year_of_capacity_data, dtype: int64
----------

12. generation_gwh_2014: 372
2431.82359    509
0.00000        28
6803.31250      1
4735.13000      1
145.81400       1
              ...
6224.00000      1
268.48085       1
1255.73200      1
164.32425       1
1153.65300      1
Name: generation_gwh_2014, Length: 372, dtype: int64
----------

13. generation_gwh_2015: 397
2428.226946    485
0.000000        27
2985.139300      1
8076.810500      1
1.093950         1
               ...
665.197300       1
1516.360100      1
```

```
741.862050       1
183.298900       1
7130.507000      1
Name: generation_gwh_2015, Length: 397, dtype: int64
----------

14. generation_gwh_2016: 404
2467.936859    473
0.000000        30
8470.570000      2
1511.000000      2
226.969450       1
              ...
433.848000       1
283.748110       1
259.943750       1
403.960000       1
307.872900       1
Name: generation_gwh_2016, Length: 404, dtype: int64
----------

15. generation_gwh_2017: 409
2547.759305    467
0.000000        32
170.085300       2
15.611550        1
549.869300       1
              ...
214.482200       1
272.739450       1
2887.000000      1
12.736000        1
158.732350       1
Name: generation_gwh_2017, Length: 409, dtype: int64
----------

16. generation_gwh_2018: 411
2600.804099    459
0.000000        39
881.020000       1
122.365100       1
805.482350       1
              ...
833.700550       1
980.254100       1
33.889700        1
6474.614250      1
192.015100       1
Name: generation_gwh_2018, Length: 411, dtype: int64
----------

17. generation_data_source: 1
Central Electricity Authority    907
Name: generation_data_source, dtype: int64
----------
```

In [481…
```
from sklearn.preprocessing import LabelEncoder
```

In [482…
```
lb=LabelEncoder()
```

In [483…
```
e=lb.fit_transform(L['country_long'])
```

In [484…
```
pd.Series(e)
```

Out[484…
```
0      0
1      0
2      0
3      0
4      0
      ..
902    0
903    0
904    0
905    0
906    0
Length: 907, dtype: int32
```

In [485…

```
L['country_long']=e
L
```

Out[485…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	source	geolocation_source	year_c
0	0	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	Oil	National Renewable Energy Laboratory	National Renewable Energy Laboratory	
1	0	ADITYA CEMENT WORKS	WRI1019881	98.0	24.7663	74.6090	Coal	Oil	Ultratech Cement ltd	WRI	
2	0	AES Saurashtra Windfarms	WRI1026669	39.2	21.9038	69.3732	Wind	Oil	CDM	WRI	
3	0	AGARTALA GT	IND0000001	135.0	23.8712	91.3602	Gas	Oil	Central Electricity Authority	WRI	
4	0	AKALTARA TPP	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	Central Electricity Authority	WRI	
...	...	...	...	...	...	...	...	...	...	...	
902	0	YERMARUS TPP	IND0000513	1600.0	16.2949	77.3568	Coal	Oil	Central Electricity Authority	WRI	
903	0	Yelesandra Solar Power Plant	WRI1026222	3.0	12.8932	78.1654	Solar	Oil	Karnataka Power Corporation Limited	Industry About	
904	0	Yelisirur wind power project	WRI1026776	25.5	15.2758	75.5811	Wind	Oil	CDM	WRI	
905	0	ZAWAR MINES	WRI1019901	80.0	24.3500	73.7477	Coal	Oil	Hindustan Zinc ltd	WRI	
906	0	iEnergy Theni Wind Farm	WRI1026761	16.5	9.9344	77.4768	Wind	Oil	CDM	WRI	

907 rows × 17 columns

In [486…

```
e_=lb.fit_transform(L['name'])
pd.Series(e_)
L['name']=e_
L
```

Out[486…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	source	geolocation_source	year_of_capa
0	0	0	WRI1020239	2.5	28.1839	73.2407	Solar	Oil	National Renewable Energy Laboratory	National Renewable Energy Laboratory	
1	0	1	WRI1019881	98.0	24.7663	74.6090	Coal	Oil	Ultratech Cement ltd	WRI	
2	0	2	WRI1026669	39.2	21.9038	69.3732	Wind	Oil	CDM	WRI	
3	0	3	IND0000001	135.0	23.8712	91.3602	Gas	Oil	Central Electricity Authority	WRI	
4	0	4	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	Central Electricity Authority	WRI	
...	...	...	...	...	...	...	...	...	...	...	
902	0	902	IND0000513	1600.0	16.2949	77.3568	Coal	Oil	Central Electricity Authority	WRI	
903	0	903	WRI1026222	3.0	12.8932	78.1654	Solar	Oil	Karnataka Power Corporation Limited	Industry About	
904	0	904	WRI1026776	25.5	15.2758	75.5811	Wind	Oil	CDM	WRI	
905	0	905	WRI1019901	80.0	24.3500	73.7477	Coal	Oil	Hindustan Zinc ltd	WRI	

906	0	906	WRI1026761	16.5	9.9344	77.4768	Wind	Oil	CDM	WRI	

907 rows × 17 columns

In [487…

```
f=lb.fit_transform(L['gppd_idnr'])
pd.Series(f)
L['gppd_idnr']=f
L
```

Out[487…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	source	geolocation_source	year_of_capac
0	0	0	657	2.5	28.1839	73.2407	Solar	Oil	National Renewable Energy Laboratory	National Renewable Energy Laboratory	
1	0	1	519	98.0	24.7663	74.6090	Coal	Oil	Ultratech Cement ltd	WRI	
2	0	2	853	39.2	21.9038	69.3732	Wind	Oil	CDM	WRI	
3	0	3	0	135.0	23.8712	91.3602	Gas	Oil	Central Electricity Authority	WRI	
4	0	4	1	1800.0	21.9603	82.4091	Coal	Oil	Central Electricity Authority	WRI	
...	...	...	...	...	...	...	...	...	...	...	
902	0	902	491	1600.0	16.2949	77.3568	Coal	Oil	Central Electricity Authority	WRI	
903	0	903	822	3.0	12.8932	78.1654	Solar	Oil	Karnataka Power Corporation Limited	Industry About	
904	0	904	891	25.5	15.2758	75.5811	Wind	Oil	CDM	WRI	
905	0	905	539	80.0	24.3500	73.7477	Coal	Oil	Hindustan Zinc ltd	WRI	
906	0	906	876	16.5	9.9344	77.4768	Wind	Oil	CDM	WRI	

907 rows × 17 columns

In [488…

```
e_=lb.fit_transform(L['primary_fuel'])
pd.Series(e_)
L['primary_fuel']=e_
L
```

Out[488…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	source	geolocation_source	year_of_capac
0	0	0	657	2.5	28.1839	73.2407	6	Oil	National Renewable Energy Laboratory	National Renewable Energy Laboratory	
1	0	1	519	98.0	24.7663	74.6090	1	Oil	Ultratech Cement ltd	WRI	
2	0	2	853	39.2	21.9038	69.3732	7	Oil	CDM	WRI	
3	0	3	0	135.0	23.8712	91.3602	2	Oil	Central Electricity Authority	WRI	
4	0	4	1	1800.0	21.9603	82.4091	1	Oil	Central Electricity Authority	WRI	
...	...	...	...	...	...	...	...	...	...	...	
902	0	902	491	1600.0	16.2949	77.3568	1	Oil	Central Electricity Authority	WRI	
903	0	903	822	3.0	12.8932	78.1654	6	Oil	Karnataka Power Corporation Limited	Industry About	
904	0	904	891	25.5	15.2758	75.5811	7	Oil	CDM	WRI	

905	0	905	539	80.0	24.3500	73.7477	1	Oil	Hindustan Zinc ltd	WRI	
906	0	906	876	16.5	9.9344	77.4768	7	Oil	CDM	WRI	

907 rows × 17 columns

In [489…

```
e_=lb.fit_transform(L['other_fuel1'])
pd.Series(e_)
L['other_fuel1']=e_
L
```

Out[489…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	source	geolocation_source	year_of_capac
0	0	0	657	2.5	28.1839	73.2407	6	2	National Renewable Energy Laboratory	National Renewable Energy Laboratory	
1	0	1	519	98.0	24.7663	74.6090	1	2	Ultratech Cement ltd	WRI	
2	0	2	853	39.2	21.9038	69.3732	7	2	CDM	WRI	
3	0	3	0	135.0	23.8712	91.3602	2	2	Central Electricity Authority	WRI	
4	0	4	1	1800.0	21.9603	82.4091	1	2	Central Electricity Authority	WRI	
...	...	...	...	...	...	...	...	...	...	...	
902	0	902	491	1600.0	16.2949	77.3568	1	2	Central Electricity Authority	WRI	
903	0	903	822	3.0	12.8932	78.1654	6	2	Karnataka Power Corporation Limited	Industry About	
904	0	904	891	25.5	15.2758	75.5811	7	2	CDM	WRI	
905	0	905	539	80.0	24.3500	73.7477	1	2	Hindustan Zinc ltd	WRI	
906	0	906	876	16.5	9.9344	77.4768	7	2	CDM	WRI	

907 rows × 17 columns

In [490…

```
e_=lb.fit_transform(L['source'])
pd.Series(e_)
L['source']=e_
L
```

Out[490…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	source	geolocation_source	year_of_capacity_
0	0	0	657	2.5	28.1839	73.2407	6	2	109	National Renewable Energy Laboratory	20
1	0	1	519	98.0	24.7663	74.6090	1	2	174	WRI	20
2	0	2	853	39.2	21.9038	69.3732	7	2	21	WRI	20
3	0	3	0	135.0	23.8712	91.3602	2	2	22	WRI	20
4	0	4	1	1800.0	21.9603	82.4091	1	2	22	WRI	20
...	...	...	...	...	...	...	...	...	...	...	
902	0	902	491	1600.0	16.2949	77.3568	1	2	22	WRI	20
903	0	903	822	3.0	12.8932	78.1654	6	2	77	Industry About	20
904	0	904	891	25.5	15.2758	75.5811	7	2	21	WRI	20
905	0	905	539	80.0	24.3500	73.7477	1	2	59	WRI	20

906	0	906	876	16.5	9.9344	77.4768	7	2	21	WRI	20

907 rows × 17 columns

In [491…

```
e_=lb.fit_transform(L['geolocation_source'])
pd.Series(e_)
L['geolocation_source']=e_
L
```

Out[491…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	source	geolocation_source	year_of_capacity_
0	0	0	657	2.5	28.1839	73.2407	6	2	109	1	20
1	0	1	519	98.0	24.7663	74.6090	1	2	174	2	20
2	0	2	853	39.2	21.9038	69.3732	7	2	21	2	20
3	0	3	0	135.0	23.8712	91.3602	2	2	22	2	20
4	0	4	1	1800.0	21.9603	82.4091	1	2	22	2	20
...	...	...	...	...	...	...	...	...	...	...	
902	0	902	491	1600.0	16.2949	77.3568	1	2	22	2	20
903	0	903	822	3.0	12.8932	78.1654	6	2	77	0	20
904	0	904	891	25.5	15.2758	75.5811	7	2	21	2	20
905	0	905	539	80.0	24.3500	73.7477	1	2	59	2	20
906	0	906	876	16.5	9.9344	77.4768	7	2	21	2	20

907 rows × 17 columns

In [492…

```
e_=lb.fit_transform(L['generation_data_source'])
pd.Series(e_)
L['generation_data_source']=e_
L
```

Out[492…

	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	source	geolocation_source	year_of_capacity_
0	0	0	657	2.5	28.1839	73.2407	6	2	109	1	20
1	0	1	519	98.0	24.7663	74.6090	1	2	174	2	20
2	0	2	853	39.2	21.9038	69.3732	7	2	21	2	20
3	0	3	0	135.0	23.8712	91.3602	2	2	22	2	20
4	0	4	1	1800.0	21.9603	82.4091	1	2	22	2	20
...	...	...	...	...	...	...	...	...	...	...	
902	0	902	491	1600.0	16.2949	77.3568	1	2	22	2	20
903	0	903	822	3.0	12.8932	78.1654	6	2	77	0	20
904	0	904	891	25.5	15.2758	75.5811	7	2	21	2	20
905	0	905	539	80.0	24.3500	73.7477	1	2	59	2	20
906	0	906	876	16.5	9.9344	77.4768	7	2	21	2	20

907 rows × 17 columns

In [493…

```
L=L.drop(['latitude','longitude'],axis=1)
```

In [494…

```
L
```

Out[494…

	country_long	name	gppd_idnr	capacity_mw	primary_fuel	other_fuel1	source	geolocation_source	year_of_capacity_data	generation_gwl
0	0	0	657	2.5	6	2	109	1	2019.0	2431.8
1	0	1	519	98.0	1	2	174	2	2019.0	2431.8

2	0	2	853	39.2	7	2	21	2	2019.0	2431.8
3	0	3	0	135.0	2	2	22	2	2019.0	617.7
4	0	4	1	1800.0	1	2	22	2	2019.0	3035.5
...	...	...	...	...	...	...	...	...	...	
902	0	902	491	1600.0	1	2	22	2	2019.0	2431.8
903	0	903	822	3.0	6	2	77	0	2019.0	2431.8
904	0	904	891	25.5	7	2	21	2	2019.0	2431.8
905	0	905	539	80.0	1	2	59	2	2019.0	2431.8
906	0	906	876	16.5	7	2	21	2	2019.0	2431.8

907 rows × 15 columns

In [495…

```
L
```

Out[495…

	country_long	name	gppd_idnr	capacity_mw	primary_fuel	other_fuel1	source	geolocation_source	year_of_capacity_data	generation_gwh
0	0	0	657	2.5	6	2	109	1	2019.0	2431.8
1	0	1	519	98.0	1	2	174	2	2019.0	2431.8
2	0	2	853	39.2	7	2	21	2	2019.0	2431.8
3	0	3	0	135.0	2	2	22	2	2019.0	617.7
4	0	4	1	1800.0	1	2	22	2	2019.0	3035.5
...	...	...	...	...	...	...	...	...	...	
902	0	902	491	1600.0	1	2	22	2	2019.0	2431.8
903	0	903	822	3.0	6	2	77	0	2019.0	2431.8
904	0	904	891	25.5	7	2	21	2	2019.0	2431.8
905	0	905	539	80.0	1	2	59	2	2019.0	2431.8
906	0	906	876	16.5	7	2	21	2	2019.0	2431.8

907 rows × 15 columns

In [496…

```
## checking correlation
```

In [497…

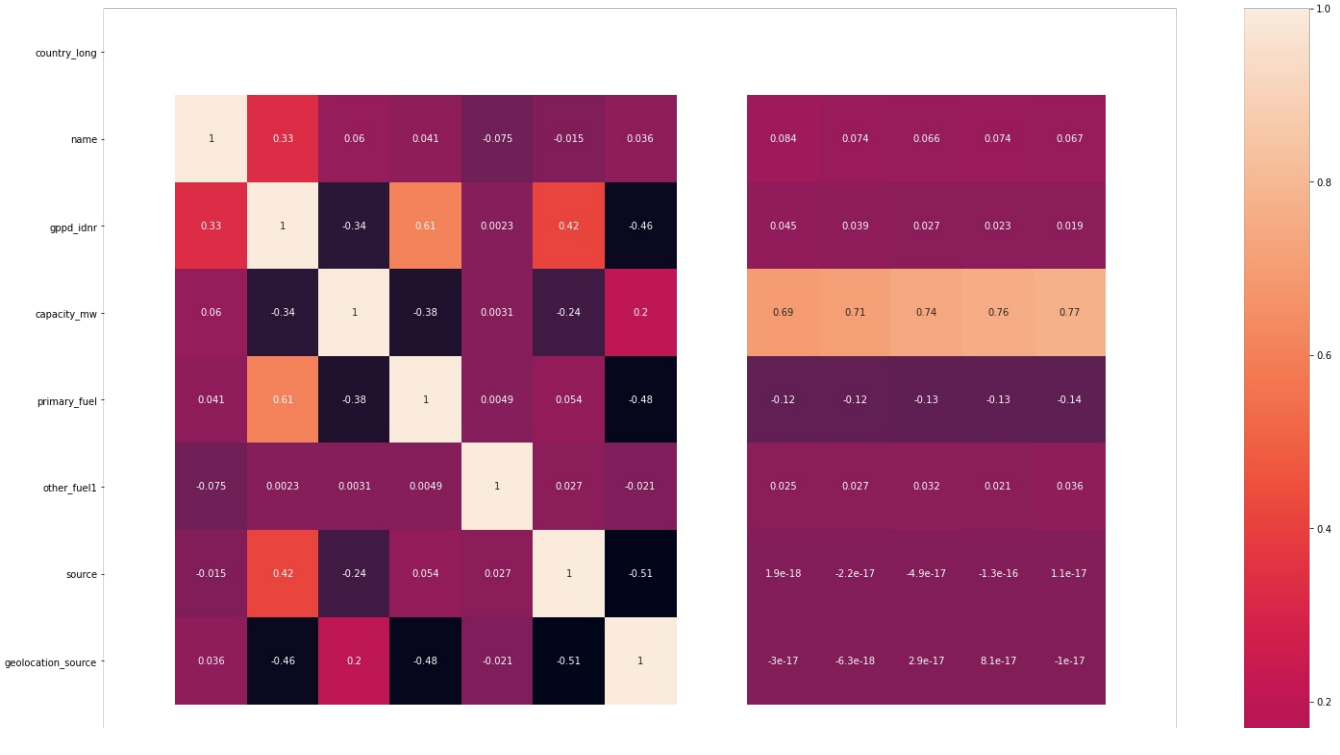
```
corr=L.corr()
```

In [498…

```
sns.heatmap(corr,annot=True)
```

Out[498…

```
<AxesSubplot:>
```



In [499…

```
L=L.drop(['generation_gwh_2016','generation_gwh_2017','generation_gwh_2018'],axis=1)
L
```

Out[499…

	country_long	name	gppd_idnr	capacity_mw	primary_fuel	other_fuel1	source	geolocation_source	year_of_capacity_data	generation_gwh
0	0	0	657	2.5	6	2	109	1	2019.0	2431.8
1	0	1	519	98.0	1	2	174	2	2019.0	2431.8
2	0	2	853	39.2	7	2	21	2	2019.0	2431.8
3	0	3	0	135.0	2	2	22	2	2019.0	617.7
4	0	4	1	1800.0	1	2	22	2	2019.0	3035.5
...	...	...	...	...	...	...	...	...	...	
902	0	902	491	1600.0	1	2	22	2	2019.0	2431.8
903	0	903	822	3.0	6	2	77	0	2019.0	2431.8
904	0	904	891	25.5	7	2	21	2	2019.0	2431.8
905	0	905	539	80.0	1	2	59	2	2019.0	2431.8
906	0	906	876	16.5	7	2	21	2	2019.0	2431.8

907 rows × 12 columns

In [500…

```
x=L.drop(['primary_fuel'],axis=1)
x
```

Out[500…

	country_long	name	gppd_idnr	capacity_mw	other_fuel1	source	geolocation_source	year_of_capacity_data	generation_gwh_2014	gener
0	0	0	657	2.5	2	109	1	2019.0	2431.823590	
1	0	1	519	98.0	2	174	2	2019.0	2431.823590	
2	0	2	853	39.2	2	21	2	2019.0	2431.823590	
3	0	3	0	135.0	2	22	2	2019.0	617.789264	
4	0	4	1	1800.0	2	22	2	2019.0	3035.550000	
...	...	...	...	...	...	...	...	...	...	
902	0	902	491	1600.0	2	22	2	2019.0	2431.823590	
903	0	903	822	3.0	2	77	0	2019.0	2431.823590	
904	0	904	891	25.5	2	21	2	2019.0	2431.823590	
905	0	905	539	80.0	2	59	2	2019.0	2431.823590	
906	0	906	876	16.5	2	21	2	2019.0	2431.823590	

907 rows × 11 columns

In [501…
```
y=L.primary_fuel
y
```

Out[501…
```
0      6
1      1
2      7
3      2
4      1
      ..
902    1
903    6
904    7
905    1
906    7
Name: primary_fuel, Length: 907, dtype: int32
```

In [502…
```
from sklearn.preprocessing import StandardScaler
```

In [503…
```
st=StandardScaler()
```

In [504…
```
st.fit_transform(x)
```

Out[504…
```
array([[ 0.00000000e+00, -1.73014221e+00,  7.79136890e-01, ...,
         1.70709359e-16,  0.00000000e+00,  0.00000000e+00],
       [ 0.00000000e+00, -1.72632291e+00,  2.52073700e-01, ...,
         1.70709359e-16,  0.00000000e+00,  0.00000000e+00],
       [ 0.00000000e+00, -1.72250361e+00,  1.52771939e+00, ...,
         1.70709359e-16,  0.00000000e+00,  0.00000000e+00],
       ...,
       [ 0.00000000e+00,  1.72250361e+00,  1.67285273e+00, ...,
         1.70709359e-16,  0.00000000e+00,  0.00000000e+00],
       [ 0.00000000e+00,  1.72632291e+00,  3.28459669e-01, ...,
         1.70709359e-16,  0.00000000e+00,  0.00000000e+00],
       [ 0.00000000e+00,  1.73014221e+00,  1.61556326e+00, ...,
         1.70709359e-16,  0.00000000e+00,  0.00000000e+00]])
```

In [505…
```
from sklearn.model_selection import train_test_split,cross_val_score
#importing models
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier,AdaBoostClassifier,GradientBoostingClassifier
```

In [506…
```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.30,random_state=41)
```

In [507…
```
kn=KNeighborsClassifier()
```

In [508…
```
kn.fit(x_train,y_train)
```

Out[508…
```
KNeighborsClassifier()
```

In [509…
```
y_pred=kn.predict(x_test)
```

In [510…
```
from sklearn.metrics import accuracy_score,confusion_matrix,classification_report
```

In [511…
```
accuracy_score(y_test,y_pred)
```

Out[511…
```
0.7142857142857143
```

In [512…
```
confusion_matrix(y_test,y_pred)
```

Out[512…
```
array([[ 8,  7,  0,  0,  0,  0,  0,  0],
       [ 9, 63,  3, 10,  0,  0,  0,  1],
       [ 1,  4,  6,  8,  0,  0,  0,  0],
       [ 0, 13,  2, 60,  0,  0,  0,  0],
       [ 0,  2,  0,  0,  0,  0,  0,  0],
       [ 2,  3,  1,  5,  0,  0,  0,  0],
       [ 0,  1,  0,  0,  0,  0, 29,  2],
       [ 0,  3,  0,  0,  0,  0,  1, 29]], dtype=int64)
```

In [513…
```
classification_report(y_test,y_pred)
```

Out[513…
```
'              precision    recall  f1-score   support\n\n           0       0.40      0.53      0.46        15\n
1       0.66      0.73      0.69        86\n           2       0.50      0.32      0.39        19\n           3
0.72      0.80      0.76        75\n           4       0.00      0.00      0.00         2\n           5       0.0
0      0.00      0.00        11\n           6       0.97      0.91      0.94        32\n           7       0.91
0.88      0.89        33\n\n    accuracy                           0.71       273\n   macro avg       0.52      0
.52      0.52       273\nweighted avg       0.68      0.71      0.70       273\n'
```

In [514…
```
## svc
```

In [515…
```
sv=SVC()
```

In [516…
```
sv.fit(x_train,y_train)
```

Out[516…
```
SVC()
```

In [517…
```
y_pred=sv.predict(x_test)
```

In [518…
```
accuracy_score(y_test,y_pred)
```

Out[518…
```
0.5494505494505495
```

In [519…
```
confusion_matrix(y_test,y_pred)
```

Out[519…
```
array([[ 0,  0,  0,  0,  0,  0, 15,  0],
       [ 0, 55,  0, 12,  0,  0, 19,  0],
       [ 0,  2,  0, 13,  0,  0,  4,  0],
       [ 0,  9,  0, 63,  0,  0,  3,  0],
       [ 0,  2,  0,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  8,  0,  0,  3,  0],
       [ 0,  0,  0,  0,  0,  0, 32,  0],
       [ 0,  0,  0,  0,  0,  0, 33,  0]], dtype=int64)
```

In [520…
```
classification_report(y_test,y_pred)
```

Out[520…
```
'              precision    recall  f1-score   support\n\n           0       0.00      0.00      0.00        15\n
1       0.81      0.64      0.71        86\n           2       0.00      0.00      0.00        19\n           3
0.66      0.84      0.74        75\n           4       0.00      0.00      0.00         2\n           5       0.0
0      0.00      0.00        11\n           6       0.29      1.00      0.45        32\n           7       0.00
0.00      0.00        33\n\n    accuracy                           0.55       273\n   macro avg       0.22      0
.31      0.24       273\nweighted avg       0.47      0.55      0.48       273\n'
```

In [521…
```
##Decisontreeclassifier
```

In [522…
```
dt=DecisionTreeClassifier()
```

In [523…
```
dt.fit(x_train,y_train)
```

Out[523… DecisionTreeClassifier()

In [524…
```
y_pred=dt.predict(x_test)
```

In [525…
```
accuracy_score(y_test,y_pred)
```

Out[525… 0.6776556776556777

In [526…
```
confusion_matrix(y_test,y_pred)
```

Out[526…
```
array([[ 6,  8,  0,  0,  0,  1,  0,  0],
       [ 3, 54,  8, 12,  4,  5,  0,  0],
       [ 1,  6,  3,  9,  0,  0,  0,  0],
       [ 0,  9,  9, 55,  0,  2,  0,  0],
       [ 0,  0,  0,  0,  1,  0,  0,  1],
       [ 1,  4,  1,  3,  0,  2,  0,  0],
       [ 0,  0,  1,  0,  0,  0, 31,  0],
       [ 0,  0,  0,  0,  0,  0,  0, 33]], dtype=int64)
```

In [527…
```
classification_report(y_test,y_pred)
```

Out[527…
```
'              precision    recall  f1-score   support\n\n           0       0.55      0.40      0.46        15\n
1       0.67      0.63      0.65        86\n           2       0.14      0.16      0.15        19\n           3
0.70      0.73      0.71        75\n           4       0.20      0.50      0.29         2\n           5       0.2
0      0.18      0.19        11\n           6       1.00      0.97      0.98        32\n           7       0.97
1.00      0.99        33\n\n    accuracy                           0.68       273\n   macro avg       0.55      0
.57      0.55       273\nweighted avg       0.68      0.68      0.68       273\n'
```

In [528…
```
## random forest classifier
```

In [529…
```
rf=RandomForestClassifier()
```

In [530…
```
rf.fit(x_train,y_train)
```

Out[530… RandomForestClassifier()

In [531…
```
y_pred=rf.predict(x_test)
```

In [532…
```
accuracy_score(y_test,y_pred)
```

Out[532… 0.7509157509157509

In [533…
```
confusion_matrix(y_test,y_pred)
```

Out[533…
```
array([[10,  5,  0,  0,  0,  0,  0,  0],
       [ 3, 67,  5,  9,  0,  2,  0,  0],
       [ 1,  6,  3,  9,  0,  0,  0,  0],
       [ 0, 10,  6, 59,  0,  0,  0,  0],
       [ 0,  1,  0,  0,  0,  0,  0,  1],
       [ 0,  4,  3,  2,  0,  2,  0,  0],
       [ 0,  1,  0,  0,  0,  0, 31,  0],
       [ 0,  0,  0,  0,  0,  0,  0, 33]], dtype=int64)
```

In [534…
```
classification_report(y_test,y_pred)
```

Out[534…
```
'              precision    recall  f1-score   support\n\n           0       0.71      0.67      0.69        15\n
```

```
1       0.71      0.78      0.74        86\n           2       0.18      0.16      0.17        19\n           3
0.75      0.79      0.77        75\n           4       0.00      0.00      0.00         2\n           5       0.5
0      0.18      0.27        11\n           6       1.00      0.97      0.98        32\n           7       0.97
1.00      0.99        33\n\n    accuracy                           0.75       273\n   macro avg       0.60      0
.57      0.58       273\nweighted avg       0.74      0.75      0.74       273\n'
```

In [535…
```
## gradientBoostingClassifier
```

In [536…
```
gb=GradientBoostingClassifier()
```

In [537…
```
gb.fit(x_train,y_train)
```

Out[537… GradientBoostingClassifier()

In [538…
```
y_pred=gb.predict(x_test)
```

In [539…
```
accuracy_score(y_test,y_pred)
```

Out[539… 0.7142857142857143

In [540…
```
confusion_matrix(y_test,y_pred)
```

Out[540…
```
array([[ 9,  5,  0,  0,  0,  1,  0,  0],
       [ 2, 65,  4, 13,  1,  1,  0,  0],
       [ 1,  7,  3,  8,  0,  0,  0,  0],
       [ 0, 13,  9, 53,  0,  0,  0,  0],
       [ 0,  2,  0,  0,  0,  0,  0,  0],
       [ 1,  4,  2,  3,  0,  1,  0,  0],
       [ 0,  1,  0,  0,  0,  0, 31,  0],
       [ 0,  0,  0,  0,  0,  0,  0, 33]], dtype=int64)
```

In [541…
```
classification_report(y_test,y_pred)
```

Out[541…
```
'              precision    recall  f1-score   support\n\n           0       0.69      0.60      0.64        15\n
1       0.67      0.76      0.71        86\n           2       0.17      0.16      0.16        19\n           3
0.69      0.71      0.70        75\n           4       0.00      0.00      0.00         2\n           5       0.3
3      0.09      0.14        11\n           6       1.00      0.97      0.98        32\n           7       1.00
1.00      1.00        33\n\n    accuracy                           0.71       273\n   macro avg       0.57      0
.54      0.54       273\nweighted avg       0.70      0.71      0.70       273\n'
```

In [542…
```
## ada boost classifier
```

In [543…
```
ad=AdaBoostClassifier()
```

In [544…
```
ad.fit(x_train,y_train)
```

Out[544… AdaBoostClassifier()

In [545…
```
y_pred=ad.predict(x_test)
```

In [546…
```
accuracy_score(y_test,y_pred)
```

Out[546… 0.45054945054945056

In [547…
```
confusion_matrix(y_test,y_pred)
```

```
confusion_matrix(y_test,y_pred)
```

Out[547…
```
array([[ 5, 10,  0,  0,  0,  0,  0,  0],
       [ 5, 10,  0, 71,  0,  0,  0,  0],
       [ 1,  3,  0, 15,  0,  0,  0,  0],
       [ 0,  0,  0, 75,  0,  0,  0,  0],
       [ 0,  0,  0,  1,  0,  0,  0,  1],
       [ 0,  3,  0,  8,  0,  0,  0,  0],
       [ 8,  1,  0,  0,  0,  0,  0, 23],
       [ 0,  0,  0,  0,  0,  0,  0, 33]], dtype=int64)
```

In [548…
```
classification_report(y_test,y_pred)
```

Out[548…
```
'              precision    recall  f1-score   support\n\n           0       0.26      0.33      0.29        15\n
1       0.37      0.12      0.18        86\n           2       0.00      0.00      0.00        19\n           3
0.44      1.00      0.61        75\n           4       0.00      0.00      0.00         2\n           5       0.0
0      0.00      0.00        11\n           6       0.00      0.00      0.00        32\n           7       0.58
1.00      0.73        33\n\n    accuracy                           0.45       273\n   macro avg       0.21      0
.31      0.23       273\nweighted avg       0.32      0.45      0.33       273\n'
```

In [549…
```
## random forest classifier is working good
```

In [550…
```
## hyper parameter tuning
```

In [551…
```
params={'n_estimators':[100,200,300,400,500,600,700],'min_samples_split':[1,2,3,4],'min_samples_leaf':[1,2,3,4],'
```

In [552…
```
g=RandomizedSearchCV(RandomForestRegressor(),params,cv=10)
```

In [553…
```
g.fit(x_train,y_train)
```

Out[553…
```
RandomizedSearchCV(cv=10, estimator=RandomForestRegressor(),
                   param_distributions={'max_depth': [None, 1, 2, 3, 4, 5, 6, 7,
                                                      8],
                                        'min_samples_leaf': [1, 2, 3, 4],
                                        'min_samples_split': [1, 2, 3, 4],
                                        'n_estimators': [100, 200, 300, 400,
                                                         500, 600, 700]})
```

In [554…
```
g.best_params_
```

Out[554…
```
{'n_estimators': 100,
 'min_samples_split': 3,
 'min_samples_leaf': 3,
 'max_depth': 7}
```

In [555…
```
m=RandomForestRegressor(n_estimators= 200,
 min_samples_split= 2,
 min_samples_leaf=4,
 max_depth=7)
```

In [556…
```
m.fit(x_train,y_train)
```

Out[556…
```
RandomForestRegressor(max_depth=7, min_samples_leaf=4, n_estimators=200)
```

In [557…
```
y_pred=m.predict(x_test)
```

In [558…
```
##
```

In [559…
```
##finalizing the rf model
```

In [560…
```
a=np.array(y_test)
```

In [561…
```
predicted=np.array(rf.predict(x_test))
```

In [562…
```
df_com1=pd.DataFrame({'true':a,'pred':predicted})
```

In [563…
```
df_com1
```

Out[563…

	true	pred
0	3	3
1	1	1
2	1	1
3	5	1
4	3	3
...	...	...
268	7	7
269	3	3
270	6	6
271	2	3
272	3	3

273 rows × 2 columns

In [564…
```
## since they are multi class labels
```

In [565…
```
import pickle
```

In [566…
```
filename='GLOBAL_POWER_PLANT.pkl'
```

In [567…
```
pickle.dump(rf,open(filename,'wb'))
```

In [ ]:
```

```

In [ ]:
```

```