In [2]:

```python
import pandas as pd
from pandas.plotting import scatter_matrix
import numpy as np
from numpy import percentile
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn import svm
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import r2_score
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
```

In [3]:

```python
data =pd.read_csv("global power plant.csv")
```

In [4]:

```python
data
```

Out[4]:

	country	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	other_fuel2	...	year_of_capacity_data	generation_gwh_2013	generation_gwh_2014	generation_gwh_2015	generation_gwh_2
0	IND	India	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	NaN	NaN	...	NaN	NaN	NaN	NaN	
1	IND	India	ADITYA CEMENT WORKS	WRI1019881	98.0	24.7663	74.6090	Coal	NaN	NaN	...	NaN	NaN	NaN	NaN	
2	IND	India	AES Saurashtra Windfarms	WRI1026669	39.2	21.9038	69.3732	Wind	NaN	NaN	...	NaN	NaN	NaN	NaN	
3	IND	India	AGARTALA GT	IND0000001	135.0	23.8712	91.3602	Gas	NaN	NaN	...	2019.0	NaN	617.789264	843.747000	886.004
4	IND	India	AKALTARA TPP	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	NaN	...	2019.0	NaN	3035.550000	5916.370000	6243.000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
902	IND	India	YERMARUS TPP	IND0000513	1600.0	16.2949	77.3568	Coal	Oil	NaN	...	2019.0	NaN	NaN	0.994875	233.596
903	IND	India	Yelesandra Solar Power Plant	WRI1026222	3.0	12.8932	78.1654	Solar	NaN	NaN	...	NaN	NaN	NaN	NaN	
904	IND	India	Yelisirur wind power project	WRI1026776	25.5	15.2758	75.5811	Wind	NaN	NaN	...	NaN	NaN	NaN	NaN	
905	IND	India	ZAWAR MINES	WRI1019901	80.0	24.3500	73.7477	Coal	NaN	NaN	...	NaN	NaN	NaN	NaN	
906	IND	India	iEnergy Theni Wind Farm	WRI1026761	16.5	9.9344	77.4768	Wind	NaN	NaN	...	NaN	NaN	NaN	NaN	

907 rows × 27 columns

In [5]:

```python
data.shape
```

Out[5]:

```
(907, 27)
```

In [6]:

```python
data.columns
```

Out[6]:

```
Index(['country', 'country_long', 'name', 'gppd_idnr', 'capacity_mw',
       'latitude', 'longitude', 'primary_fuel', 'other_fuel1', 'other_fuel2',
       'other_fuel3', 'commissioning_year', 'owner', 'source', 'url',
       'geolocation_source', 'wepp_id', 'year_of_capacity_data',
       'generation_gwh_2013', 'generation_gwh_2014', 'generation_gwh_2015',
       'generation_gwh_2016', 'generation_gwh_2017', 'generation_gwh_2018',
       'generation_gwh_2019', 'generation_data_source',
       'estimated_generation_gwh'],
      dtype='object')
```

In [7]:

```python
data.isnull().sum()
```

Out[7]:

```
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 [8]:

```python
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 907 entries, 0 to 906
Data columns (total 27 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   country                   907 non-null    object
 1   country_long              907 non-null    object
 2   name                      907 non-null    object
 3   gppd_idnr                 907 non-null    object
 4   capacity_mw               907 non-null    float64
 5   latitude                  861 non-null    float64
 6   longitude                 861 non-null    float64
 7   primary_fuel              907 non-null    object
 8   other_fuel1               198 non-null    object
 9   other_fuel2               1 non-null      object
 10  other_fuel3               0 non-null      float64
 11  commissioning_year        527 non-null    float64
 12  owner                     342 non-null    object
 13  source                    907 non-null    object
 14  url                       907 non-null    object
 15  geolocation_source        888 non-null    object
 16  wepp_id                   0 non-null      float64
 17  year_of_capacity_data     519 non-null    float64
 18  generation_gwh_2013       0 non-null      float64
 19  generation_gwh_2014       398 non-null    float64
 20  generation_gwh_2015       422 non-null    float64
 21  generation_gwh_2016       434 non-null    float64
 22  generation_gwh_2017       440 non-null    float64
 23  generation_gwh_2018       448 non-null    float64
 24  generation_gwh_2019       0 non-null      float64
 25  generation_data_source    449 non-null    object
 26  estimated_generation_gwh  0 non-null      float64
dtypes: float64(15), object(12)
memory usage: 191.4+ KB
```

In [9]:

```python
data.nunique()
```

Out[9]:

```
country                     1
country_long                1
name                      907
gppd_idnr                 907
capacity_mw               361
latitude                  836
longitude                 827
primary_fuel                8
other_fuel1                 3
other_fuel2                 1
other_fuel3                 0
commissioning_year         73
owner                     280
source                    191
url                       304
geolocation_source          3
wepp_id                     0
year_of_capacity_data       1
generation_gwh_2013         0
generation_gwh_2014       371
generation_gwh_2015       396
generation_gwh_2016       403
generation_gwh_2017       408
generation_gwh_2018       410
generation_gwh_2019         0
generation_data_source      1
estimated_generation_gwh    0
dtype: int64
```

In [10]:

```python
# filling nan values
```

In [11]:

```python
data.fillna( value =0 , inplace =True)
```

In [12]:

```python
data
```

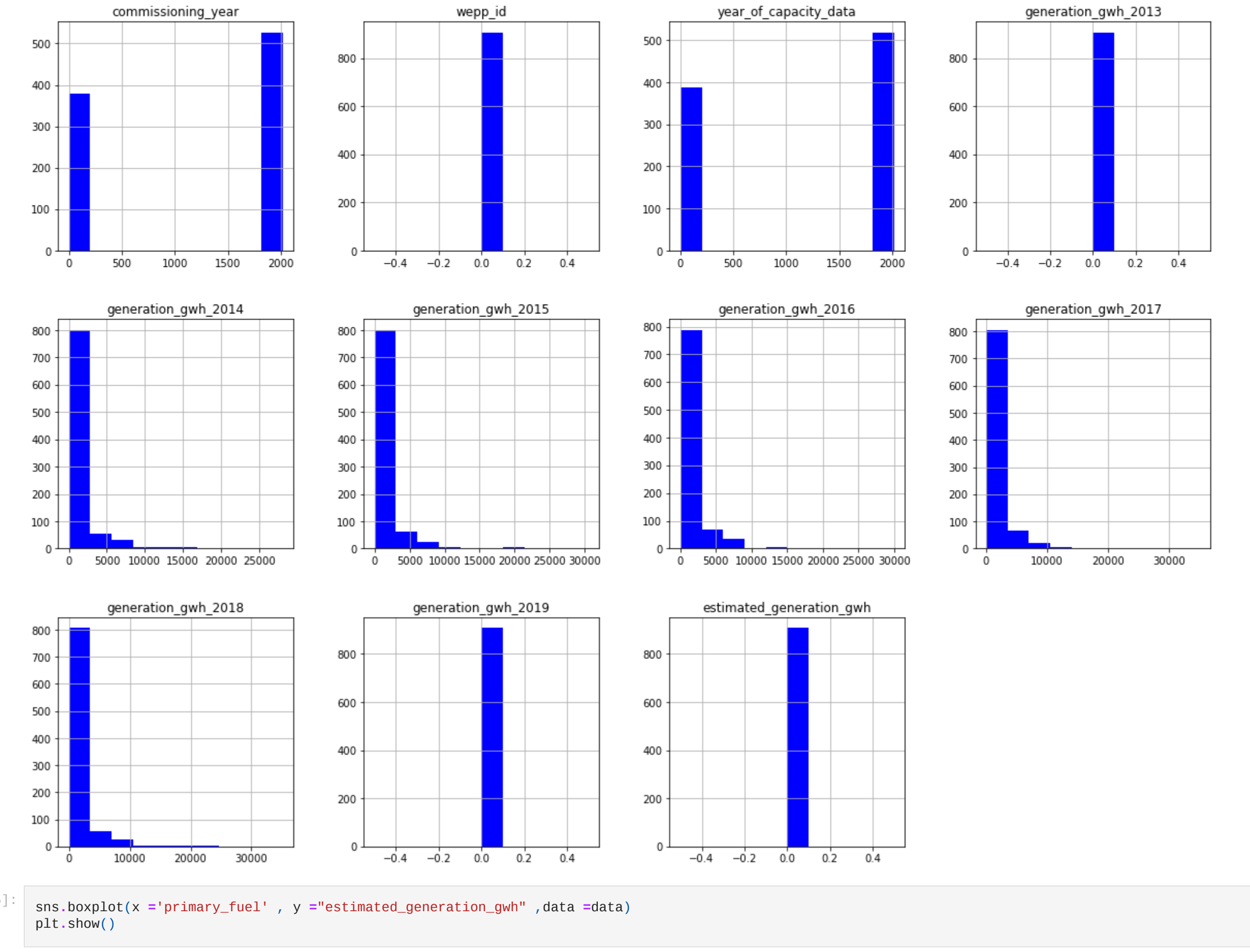
Out[12]:

	country	country_long	name	gppd_idnr	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	other_fuel2	...	year_of_capacity_data	generation_gwh_2013	generation_gwh_2014	generation_gwh_2015	generation_gwh_2
0	IND	India	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	0	0	...	0.0	0.0	0.000000	0.000000	0.000
1	IND	India	ADITYA CEMENT WORKS	WRI1019881	98.0	24.7663	74.6090	Coal	0	0	...	0.0	0.0	0.000000	0.000000	0.000
2	IND	India	AES Saurashtra Windfarms	WRI1026669	39.2	21.9038	69.3732	Wind	0	0	...	0.0	0.0	0.000000	0.000000	0.000
3	IND	India	AGARTALA GT	IND0000001	135.0	23.8712	91.3602	Gas	0	0	...	2019.0	0.0	617.789264	843.747000	886.004
4	IND	India	AKALTARA TPP	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	0	...	2019.0	0.0	3035.550000	5916.370000	6243.000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
902	IND	India	YERMARUS TPP	IND0000513	1600.0	16.2949	77.3568	Coal	Oil	0	...	2019.0	0.0	0.000000	0.994875	233.596
903	IND	India	Yelesandra Solar Power Plant	WRI1026222	3.0	12.8932	78.1654	Solar	0	0	...	0.0	0.0	0.000000	0.000000	0.000
904	IND	India	Yelisirur wind power project	WRI1026776	25.5	15.2758	75.5811	Wind	0	0	...	0.0	0.0	0.000000	0.000000	0.000
905	IND	India	ZAWAR MINES	WRI1019901	80.0	24.3500	73.7477	Coal	0	0	...	0.0	0.0	0.000000	0.000000	0.000
906	IND	India	iEnergy Theni Wind Farm	WRI1026761	16.5	9.9344	77.4768	Wind	0	0	...	0.0	0.0	0.000000	0.000000	0.000

907 rows × 27 columns

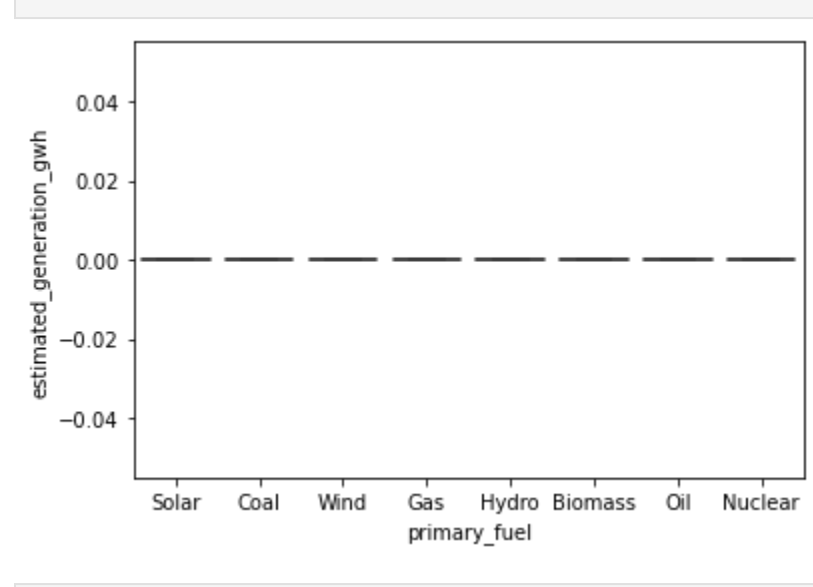
In [14]:

```python
data.hist(color ="blue" , figsize =(20,20))
plt.show()
```



In [15]:

```python
sns.boxplot(x ='primary_fuel' , y ="estimated_generation_gwh" ,data =data)
plt.show()
```



In [16]:

```python
data1 = data.drop(['country','country_long','generation_data_source','name','gppd_idnr'],axis = 1)
```

In [17]:

```python
data1.head()
```

Out[17]:

	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	other_fuel2	other_fuel3	commissioning_year	owner	source	...	wepp_id	year_of_capacity_data	generation_gwh_2013	generation_gwh_2014	generation_gwh_2015
0	2.5	28.1839	73.2407	Solar	0	0	0.0	2011.0	Solar Paces	National Renewable Energy Laboratory	...	0.0	0.0	0.0	0.000000	0.000
1	98.0	24.7663	74.6090	Coal	0	0	0.0	0.0	Ultratech Cement ltd	Ultratech Cement ltd	...	0.0	0.0	0.0	0.000000	0.000
2	39.2	21.9038	69.3732	Wind	0	0	0.0	0.0	AES	CDM	...	0.0	0.0	0.0	0.000000	0.000
3	135.0	23.8712	91.3602	Gas	0	0	0.0	2004.0	0	Central Electricity Authority	...	0.0	2019.0	0.0	617.789264	843.747
4	1800.0	21.9603	82.4091	Coal	Oil	0	0.0	2015.0	0	Central Electricity Authority	...	0.0	2019.0	0.0	3035.550000	5916.370

5 rows × 22 columns

In [18]:

```python
data1['total_generation'] = data1['generation_gwh_2013'] + data1['generation_gwh_2014'] + data1['generation_gwh_2015'] + data1['generation_gwh_2016'] + data1['generation_gwh_2017']+data1['

data2 = data1.drop(['generation_gwh_2013','generation_gwh_2014','generation_gwh_2015','generation_gwh_2016','generation_gwh_2017','generation_gwh_2018', 'generation_gwh_2019'],axis = 1)
```

In [19]:

```python
data2
```

Out[19]:

	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	other_fuel2	other_fuel3	commissioning_year	owner	source	url	geolocation_source	wepp_id	year_of_capacity_data
0	2.5	28.1839	73.2407	Solar	0	0	0.0	2011.0	Solar Paces	National Renewable Energy Laboratory	http://www.nrel.gov/csp/solarpaces/project_det...	National Renewable Energy Laboratory	0.0	0.0
1	98.0	24.7663	74.6090	Coal	0	0	0.0	0.0	Ultratech Cement ltd	Ultratech Cement ltd	http://www.ultratechcement.com/	WRI	0.0	0.0
2	39.2	21.9038	69.3732	Wind	0	0	0.0	0.0	AES	CDM	https://cdm.unfccc.int/Projects/DB/DNV-CUK1328...	WRI	0.0	0.0
3	135.0	23.8712	91.3602	Gas	0	0	0.0	2004.0	0	Central Electricity Authority	http://www.cea.nic.in/	WRI	0.0	2019.0
4	1800.0	21.9603	82.4091	Coal	Oil	0	0.0	2015.0	0	Central Electricity Authority	http://www.cea.nic.in/	WRI	0.0	2019.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
902	1600.0	16.2949	77.3568	Coal	Oil	0	0.0	2016.0	0	Central Electricity Authority	http://www.cea.nic.in/	WRI	0.0	2019.0
903	3.0	12.8932	78.1654	Solar	0	0	0.0	0.0	Karnataka Power Corporation Limited	Karnataka Power Corporation Limited	http://karnatakapower.com	Industry About	0.0	0.0
904	25.5	15.2758	75.5811	Wind	0	0	0.0	0.0	0	CDM	https://cdm.unfccc.int/Projects/DB/TUEV-RHEIN1...	WRI	0.0	0.0
905	80.0	24.3500	73.7477	Coal	0	0	0.0	0.0	Hindustan Zinc ltd	Hindustan Zinc ltd	http://www.hzlindia.com/	WRI	0.0	0.0
906	16.5	9.9344	77.4768	Wind	0	0	0.0	0.0	iEnergy Wind Farms	CDM	https://cdm.unfccc.int/Projects/DB/RWTUV134503...	WRI	0.0	0.0

907 rows × 16 columns

In [20]:

```python
data2.columns
```

Out[20]:

```
Index(['capacity_mw', 'latitude', 'longitude', 'primary_fuel', 'other_fuel1',
       'other_fuel2', 'other_fuel3', 'commissioning_year', 'owner', 'source',
       'url', 'geolocation_source', 'wepp_id', 'year_of_capacity_data',
       'estimated_generation_gwh', 'total_generation'],
      dtype='object')
```

In [21]:

```python
data2 = data2.drop(['wepp_id','url','geolocation_source'],axis = 1)
data2.head()
```

Out[21]:

	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	other_fuel2	other_fuel3	commissioning_year	owner	source	year_of_capacity_data	estimated_generation_gwh	total_generation
0	2.5	28.1839	73.2407	Solar	0	0	0.0	2011.0	Solar Paces	National Renewable Energy Laboratory	0.0	0.0	0.000000
1	98.0	24.7663	74.6090	Coal	0	0	0.0	0.0	Ultratech Cement ltd	Ultratech Cement ltd	0.0	0.0	0.000000
2	39.2	21.9038	69.3732	Wind	0	0	0.0	0.0	AES	CDM	0.0	0.0	0.000000
3	135.0	23.8712	91.3602	Gas	0	0	0.0	2004.0	0	Central Electricity Authority	2019.0	0.0	3637.554320
4	1800.0	21.9603	82.4091	Coal	Oil	0	0.0	2015.0	0	Central Electricity Authority	2019.0	0.0	27859.499736

In [22]:

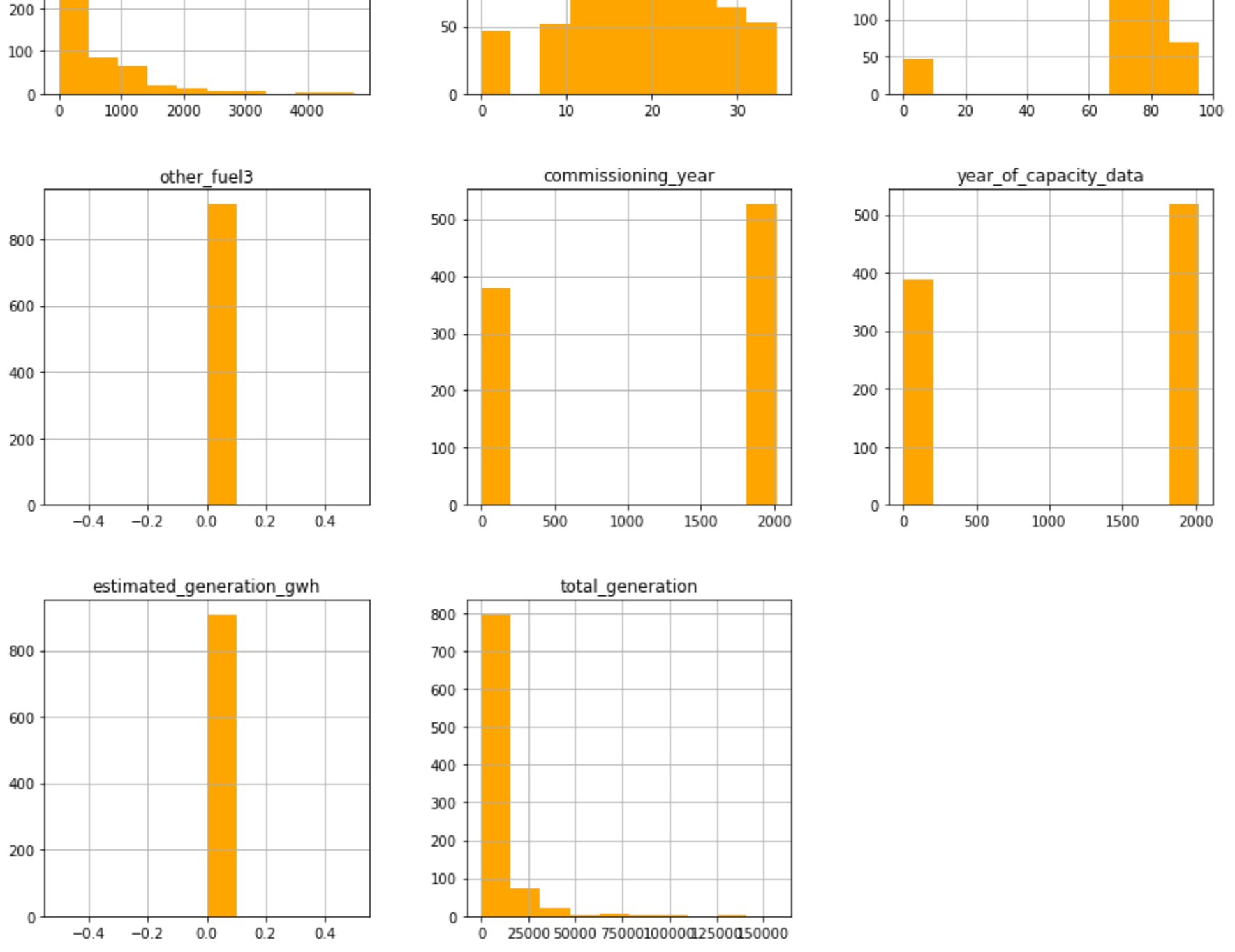
```python
data2.describe()
```

Out[22]:

	capacity_mw	latitude	longitude	other_fuel3	commissioning_year	year_of_capacity_data	estimated_generation_gwh	total_generation
count	907.000000	907.000000	907.000000	907.0	907.000000	907.000000	907.0	907.000000
mean	326.223755	20.122831	73.536147	0.0	1160.382580	1155.304300	0.0	5898.386424
std	590.085456	7.655960	17.674358	0.0	985.973139	999.466215	0.0	15335.620295
min	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000
25%	16.725000	16.172050	73.811550	0.0	0.000000	0.000000	0.0	0.000000
50%	59.200000	21.281800	76.493800	0.0	1978.000000	2019.000000	0.0	0.000000
75%	385.250000	25.176450	79.206100	0.0	2003.000000	2019.000000	0.0	3838.334000
max	4760.000000	34.649000	95.408000	0.0	2018.000000	2019.000000	0.0	156908.000000

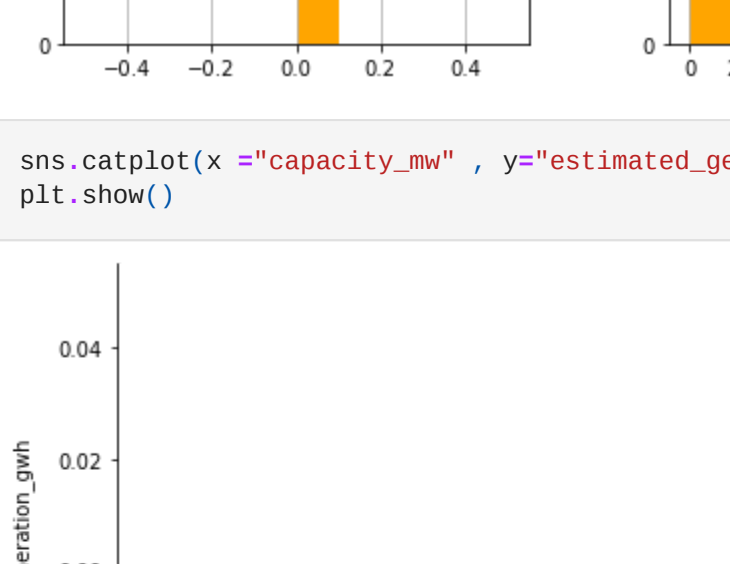
In [23]:

```python
data2.hist(color ="orange" , figsize =(15,15))
plt.show()
```



In [110…

```python
sns.catplot(x ="capacity_mw" , y="estimated_generation_gwh" ,data =data2)
plt.show()
```



In [122…

```python
#1 primary fuel
fuel_mean = data2.groupby('primary_fuel').mean()
```

In [119…

```python
fuel_mean = fuel_mean.sort_values('capacity_mw')
fuel_mean
```

Out[119…

primary_fuel	capacity_mw	latitude	longitude	other_fuel3	commissioning_year	wepp_id	year_of_capacity_data	generation_gwh_2013	generation_gwh_2014	generation_gwh_2015	generation_gwh_2016	generation_gwh_2017	generation
Biomass	20.065200	17.460458	75.679052	0.0	0.000000	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	
Solar	21.712598	23.336470	72.010522	0.0	126.826772	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	
Wind	33.429675	15.679514	65.135022	0.0	0.000000	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	
Oil	88.942000	14.715070	63.608735	0.0	1196.750000	0.0	1211.400000	0.0	71.984751	11.940547	2.638279	0.058615	
Hydro	185.026972	20.662257	73.191943	0.0	1988.709163	0.0	2019.000000	0.0	516.140858	483.699210	487.572366	503.135549	
Gas	364.818928	19.759562	77.271887	0.0	1712.565217	0.0	1726.391304	0.0	581.157629	669.692473	694.575640	715.189949	
Coal	797.826434	21.237991	77.892091	0.0	1469.527132	0.0	1479.034884	0.0	2956.209182	3189.832085	3358.699915	3530.195629	3
Nuclear	975.555556	18.081478	76.124056	0.0	1772.666667	0.0	1794.666667	0.0	3785.877017	3764.333333	3797.874444	3843.035556	3

In [129…

```python
x = fuel_mean['capacity_mw']
y = fuel_mean['estimated_generation_gwh']

print(x)
print(y)
```

```
primary_fuel
Biomass     20.065200
Coal       797.826434
Gas        364.818928
Hydro      185.026972
Nuclear    975.555556
Oil         88.942000
Solar       21.712598
Wind        33.429675
Name: capacity_mw, dtype: float64
primary_fuel
Biomass    0.0
Coal       0.0
Gas        0.0
Hydro      0.0
Nuclear    0.0
Oil        0.0
Solar      0.0
Wind       0.0
Name: estimated_generation_gwh, dtype: float64
```

In [130…

```python
re = LinearRegression()
re = reg.fit(x.values.reshape(-1,1),y)
predictions = reg.predict(x.values.reshape(-1,1))
```

In [131…

```python
r2_score(y,predictions)
```

Out[131…

```
1.0
```

In [113…

```python
# 2 capacity_mx

x = data2['capacity_mw']
y = data2['estimated_generation_gwh']

print(x)
print(y)
```

```
0         2.5
1        98.0
2        39.2
3       135.0
4      1800.0
        ...
902    1600.0
903       3.0
904      25.5
905      80.0
906      16.5
Name: capacity_mw, Length: 907, dtype: float64
0      0.0
1      0.0
2      0.0
3      0.0
4      0.0
      ...
902    0.0
903    0.0
904    0.0
905    0.0
906    0.0
Name: estimated_generation_gwh, Length: 907, dtype: float64
```

In [114…

```python
reg = LinearRegression()
reg = reg.fit(x.values.reshape(-1,1),y)
predictions = reg.predict(x.values.reshape(-1,1))
```

In [112…

```python
r2_score(y,predictions)
```

Out[112…

```
1.0
```

In [ ]: