In [22]:
```
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')
```

In [23]:
```
file =pd.read_csv("global power plant.csv")
file.head()
```

Out[23]:

	country	country_long	name	gppd_idr	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_2016
0	IND	India	ACME Solar Tower	WRI1020239	2.5	28.1839	73.2407	Solar	NaN	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	Nan
2	IND	India	AES Saurashtra Windfarms	WRI1026669	39.2	21.9038	69.3732	Wind	NaN	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	843.747	886.00443
4	IND	India	AKALTARA TPP	IND0000002	1800.0	21.9603	82.4091	Coal	Oil	NaN	...	2019.0	NaN	3035.550000	5916.370	6243.00000

5 rows × 27 columns

In [24]:
```
file.columns
```

Out[24]:
```
Index(['country', 'country_long', 'name', 'gppd_idr', '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 [25]:
```
file.shape
```

Out[25]:
```
(907, 27)
```

In [26]:
```
file.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_idr                 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 [27]:
```
file.value_counts()
```

Out[27]:
```
Series([], dtype: int64)
```

In [28]:
```
# dropping some columns

file1 =file.drop(['country','country_long','generation_data_source','name','gppd_idr','wepp_id','url','geolocation_source'],axis =1)
```

In [29]:
```
file1.head()
```

Out[29]:

	capacity_mw	latitude	longitude	primary_fuel	other_fuel1	other_fuel2	other_fuel3	commissioning_year	owner	source	year_of_capacity_data	generation_gwh_2013	generation_gwh_2014	generation_gwh_2015	generation_gwh_2
0	2.5	28.1839	73.2407	Solar	NaN	NaN	NaN	2011.0	Solar Paces	National Renewable Energy Laboratory	NaN	NaN	NaN	NaN	N
1	98.0	24.7663	74.6090	Coal	NaN	NaN	NaN	NaN	Ultratech Cement ltd	Ultratech Cement ltd	NaN	NaN	NaN	NaN	N
2	39.2	21.9038	69.3732	Wind	NaN	NaN	NaN	NaN	AES	CDM	NaN	NaN	NaN	NaN	N
3	135.0	23.8712	91.3602	Gas	NaN	NaN	NaN	2004.0	NaN	Central Electricity Authority	2019.0	NaN	617.789264	843.747	886.004
4	1800.0	21.9603	82.4091	Coal	Oil	NaN	NaN	2015.0	NaN	Central Electricity Authority	2019.0	NaN	3035.550000	5916.370	6243.000

In [30]:
```
file1.columns
```

Out[30]:
```
Index(['capacity_mw', 'latitude', 'longitude', 'primary_fuel', 'other_fuel1',
       'other_fuel2', 'other_fuel3', 'commissioning_year', 'owner', 'source',
       '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 [31]:
```
file1['total_generation'] =file1['generation_gwh_2013'] +file1['generation_gwh_2014'] +file1['generation_gwh_2015'] +file1['generation_gwh_2016'] +file1['generation_gwh_2017']+file1['gener
```

In [32]:
```
file2 = file1.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 [33]:
```
file2.head()
```

Out[33]:

	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	NaN	NaN	NaN	2011.0	Solar Paces	National Renewable Energy Laboratory	NaN	NaN	NaN
1	98.0	24.7663	74.6090	Coal	NaN	NaN	NaN	NaN	Ultratech Cement ltd	Ultratech Cement ltd	NaN	NaN	NaN
2	39.2	21.9038	69.3732	Wind	NaN	NaN	NaN	NaN	AES	CDM	NaN	NaN	NaN
3	135.0	23.8712	91.3602	Gas	NaN	NaN	NaN	2004.0	NaN	Central Electricity Authority	2019.0	NaN	NaN
4	1800.0	21.9603	82.4091	Coal	Oil	NaN	NaN	2015.0	NaN	Central Electricity Authority	2019.0	NaN	NaN

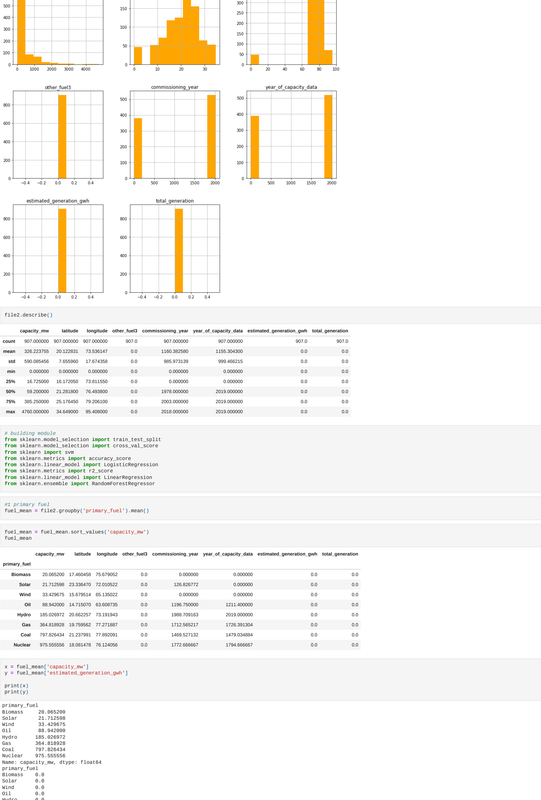
In [35]:
```
# filling values

file2.fillna( value =0 , inplace =True)
file2
```

Out[35]:

	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	NaN	NaN	0.0	2011.0	Solar Paces	National Renewable Energy Laboratory	0.0	0.0	0.0
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
2	39.2	21.9038	69.3732	Wind	0	0	0.0	0.0	AES	CDM	0.0	0.0	0.0
3	135.0	23.8712	91.3602	Gas	0	0	0.0	2004.0	0	Central Electricity Authority	2019.0	0.0	0.0
4	1800.0	21.9603	82.4091	Coal	Oil	0	0.0	2015.0	0	Central Electricity Authority	2019.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
902	1600.0	16.2949	77.3568	Coal	Oil	0	0.0	2016.0	0	Central Electricity Authority	2019.0	0.0	0.0
903	3.0	12.8932	78.1654	Solar	0	0	0.0	0.0	Karnataka Power Corporation Limited	Karnataka Power Corporation Limited	0.0	0.0	0.0
904	25.5	15.2758	75.5811	Wind	0	0	0.0	0.0	0	CDM	0.0	0.0	0.0
905	80.0	24.3500	73.7477	Coal	0	0	0.0	0.0	Hindustan Zinc ltd	Hindustan Zinc ltd	0.0	0.0	0.0
906	16.5	9.9344	77.4768	Wind	0	0	0.0	0.0	iDreegy Wind Farms	CDM	0.0	0.0	0.0

907 rows × 13 columns

In [36]:
```
file2.hist(color ="orange" , figsize =(15,15))
plt.show()
```



In [37]:
```
file2.describe()
```

Out[37]:

	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	907.0
mean	326.223755	20.123231	73.536147	0.0	1160.382580	1155.304300	0.0	0.0
std	590.085456	7.605960	17.674358	0.0	985.977139	999.496215	0.0	0.0
min	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.0
25%	16.725000	16.172050	73.811550	0.0	0.000000	0.000000	0.0	0.0
50%	59.200000	21.281800	76.493000	0.0	1978.000000	2019.000000	0.0	0.0
75%	385.250000	25.176450	79.206100	0.0	2003.000000	2019.000000	0.0	0.0
max	4760.000000	34.649000	95.408000	0.0	2018.000000	2019.000000	0.0	0.0

In [38]:
```
# building models
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 [40]:
```
#1 primary fuel
fuel_mean = file2.groupby('primary_fuel').mean()
```

In [41]:
```
fuel_mean = fuel_mean.sort_values('capacity_mw')
fuel_mean
```

Out[41]:

primary_fuel	capacity_mw	latitude	longitude	other_fuel3	commissioning_year	year_of_capacity_data	estimated_generation_gwh	total_generation
Biomass	20.065200	17.460458	75.679052	0.0	0.000000	0.000000	0.0	0.0
Solar	21.712500	23.536470	72.010522	0.0	126.824272	0.000000	0.0	0.0
Wind	33.429275	15.679514	65.135022	0.0	0.000000	0.000000	0.0	0.0
Oil	88.942000	14.715070	63.608735	0.0	1196.750000	1211.400000	0.0	0.0
Hydro	185.026972	20.662257	73.191943	0.0	1989.709163	2019.000000	0.0	0.0
Gas	364.818928	19.797662	77.271987	0.0	1712.565217	1726.391304	0.0	0.0
Coal	797.826434	21.237991	77.692091	0.0	1469.527132	1479.034884	0.0	0.0
Nuclear	975.555556	18.081478	76.124056	0.0	1772.666667	1794.666667	0.0	0.0

In [42]:
```
x = fuel_mean['capacity_mw']
y = fuel_mean['estimated_generation_gwh']

print(x)
print(y)
```

```
primary_fuel
Biomass     20.065200
Solar       21.712500
Wind        33.429275
Oil         88.942000
Hydro      185.026972
Gas        364.818928
Coal       797.826434
Nuclear    975.555556
Name: capacity_mw, dtype: float64
primary_fuel
Biomass    0.0
Solar      0.0
Wind       0.0
Oil        0.0
Hydro      0.0
Gas        0.0
Coal       0.0
Nuclear    0.0
Name: estimated_generation_gwh, dtype: float64
```

In [43]:
```
regre = LinearRegression()
regre = regre.fit(x.values.reshape(-1,1),y)
predictions = regre.predict(x.values.reshape(-1,1))
```

In [44]:
```
r2_score(y,predictions)
```

Out[44]:
```
1.0
```

In [45]:
```
# 2 capacity_mw

x = file2['capacity_mw']
y = file2['estimated_generation_gwh']

print(x)
print(y)
```

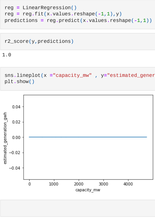
```
0          2.5
1         98.0
2         39.2
3        135.0
4       1800.0
        ...   
902     1600.0
903        3.0
904       25.5
905       80.0
906       16.5
Name: capacity_mw, Length: 907, dtype: float64
0      0.0
1      0.0
2      0.0
3      0.0
4      0.0
      ...
902    0.0
903    0.0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

In [46]:
```
reg = LinearRegression()
reg = reg.fit(x.values.reshape(-1,1),y)
predictions = reg.predict(x.values.reshape(-1,1))
```

In [47]:
```
r2_score(y,predictions)
```

Out[47]:
```
1.0
```

In [48]:
```
sns.lineplot(x ="capacity_mw" , y="estimated_generation_gwh" ,data =file2)
plt.show()
```



In [ ]:
```

```