Assignment 4

May 24, 2021

[1]:

```
import pandas as pd
import numpy as np
```

[10]:

```
df = pd.read_excel("Assignment 4/US_Solar_2019.xlsx")
df
```

[10]:

```
      Utility ID  Plant Code State  Nameplate Capacity (MW)  \
0          16572         141    AZ                      0.2
1          18454         645    FL                     19.0
2           7095         944    IL                      1.2
3          16179         960    IL                      0.3
4          14201        1172    IA                      0.8
…            …          …    …                        …
3962       62856       63800    NY                      2.0
3963       63505       63825    IL                      1.6
3964       63521       63844    NE                      5.0
3965       60293       63869    CO                      8.1
3966       56476       63928    CA                      1.5

      Summer Capacity (MW) Winter Capacity (MW)  Operating Month  \
0                      0.2                  0.2                6
1                     19.4                 19.4                2
2                      0.9                  0.9                8
3                      0.3                  0.3               11
4                      0.8                  0.1               11
…                    …                   …               …
3962                   2.0                    2                9
3963                   1.6                  1.6               12
3964                   5.0                    5               12
3965                   8.1                  8.1                7
3966                   1.4                  0.7               12

      Operating Year
0               2001
1               2017
2               2015
3               2014
```

```
4               2016
…              …
3962            2019
3963            2019
3964            2019
3965            2019
3966            2019

[3967 rows x 8 columns]
```

1 Question 1

[11]:
```
df['Operating Year'].mean()
```

[11]:
```
2015.3060247038063
```

[12]:
```
df['Operating Year'].median()
```

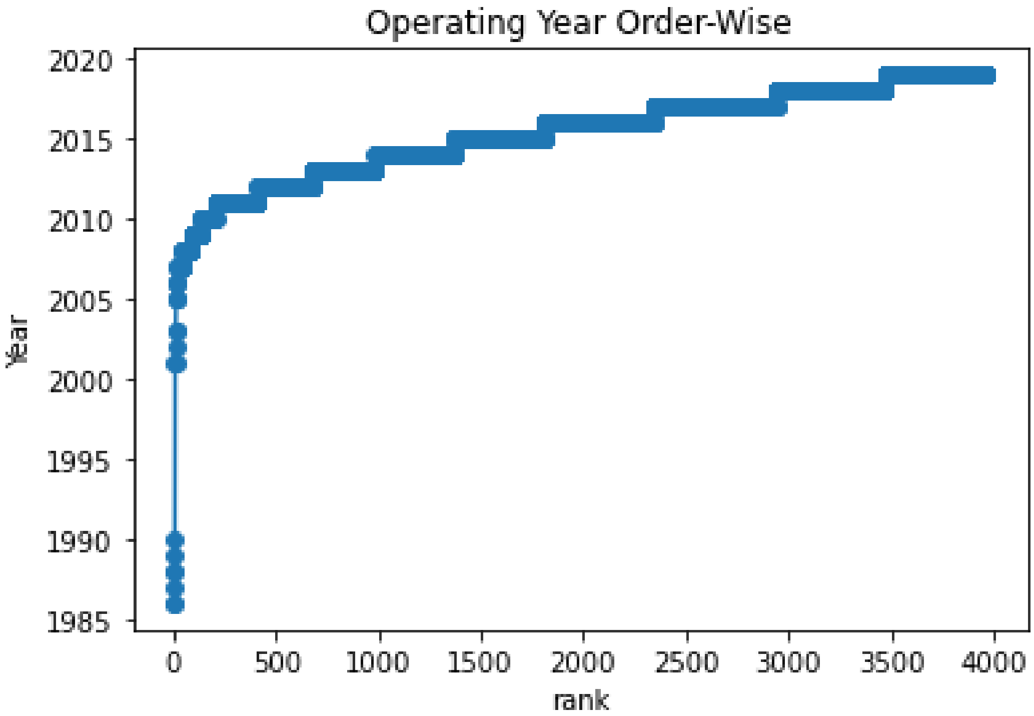
[12]:
```
2016.0
```

2 Question 2

[30]:
```
y = df['Operating Year'].sort_values().tolist()
```

[31]:
```
x = np.arange(len(y))
```

[56]:
```
import matplotlib.pyplot as plt
plt.plot(x,y, marker = 'o')
plt.xlabel('rank')
plt.ylabel('Year')
plt.title('Operating Year Order-Wise')
plt.show()
```



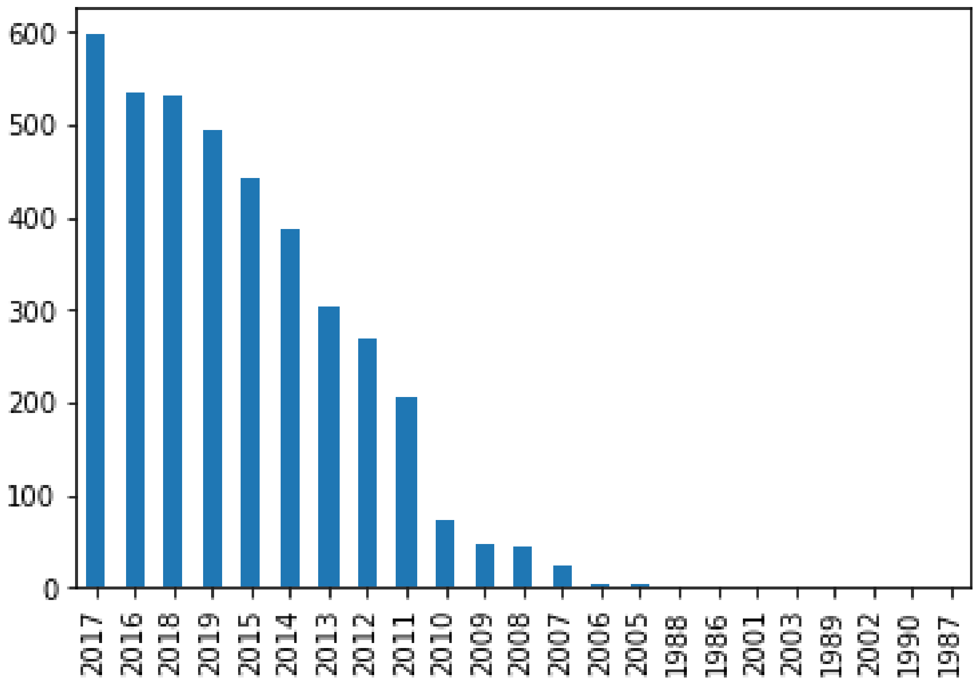
2.1

It can be seen that the median, which is at approximately 2000 rank happens to be 2015 in the diagram. Mean can also be figured to be above 2010 as the distribution is skewed towards recent operating years. Just after few first values, the ranks start corresponding to years above 2005 and it stays like this till 4000 entries.

3 Question 3

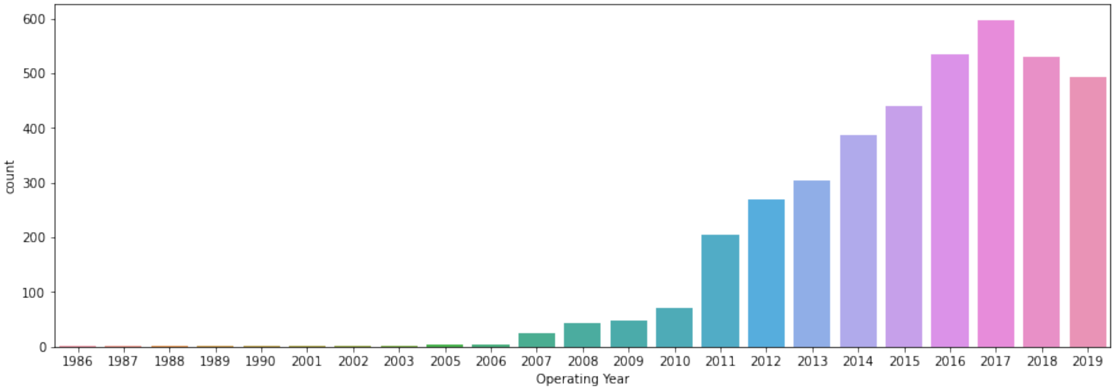
```
[87]: import seaborn as sns
      df['Operating Year'].value_counts().plot(kind ='bar')
```

```
[87]: <matplotlib.axes._subplots.AxesSubplot at 0x7f548e45f520>
```



[85]:

```
plt.figure(figsize = (15,5))
sns.countplot('Operating Year', data =df);
```



4 Question 4

4.0.1 It is the maximum possible power production of a facility when it is running on full load and utilized completely

5 Question 5

[94]:
```
print("The variance of Nameplate capacity is: " ,df[df.columns[3]].var(), "MW")
```

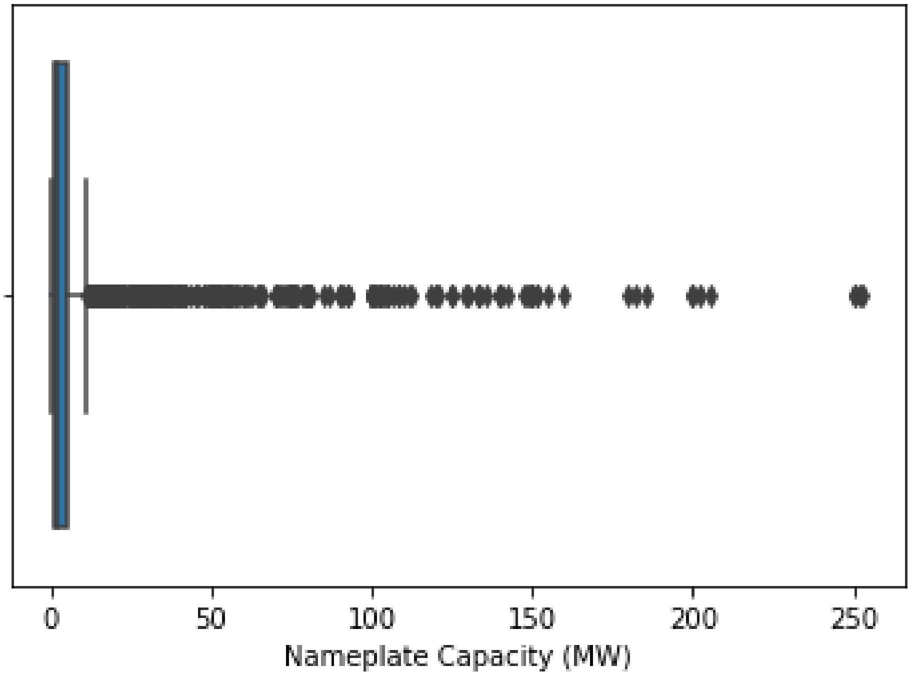
```
The variance of Nameplate capacity is:  502.97831392142785 MW
```

[97]:
```
print("The variance of Nameplate capacity is: " ,df[df.columns[3]].mean(), "MW")
```

```
The variance of Nameplate capacity is:  9.525586085202944 MW
```

[96]:
```
sns.boxplot(df.columns[3], data =df)
```

[96]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f548d6794c0>
```



6 Question 6

[107]:
```
df
```

[107]:
```
      Utility ID  Plant Code State  Nameplate Capacity (MW)  \
0          16572         141    AZ                      0.2   
1          18454         645    FL                     19.0   
2           7095         944    IL                      1.2   
3          16179         960    IL                      0.3   
4          14201        1172    IA                      0.8   
…          …         …   …                      …   
3962       62856       63800    NY                      2.0   
3963       63505       63825    IL                      1.6   
3964       63521       63844    NE                      5.0   
3965       60293       63869    CO                      8.1   
3966       56476       63928    CA                      1.5   

      Summer Capacity (MW) Winter Capacity (MW)  Operating Month  \
0                      0.2                  0.2                6   
1                     19.4                 19.4                2   
2                      0.9                  0.9                8   
3                      0.3                  0.3               11   
4                      0.8                  0.1               11   
…                    …                  …              …   
3962                   2.0                    2                9   
3963                   1.6                  1.6               12   
3964                   5.0                    5               12   
3965                   8.1                  8.1                7   
3966                   1.4                  0.7               12   

      Operating Year  
0               2001  
1               2017  
2               2015  
3               2014  
4               2016  
…             …  
3962            2019  
3963            2019  
3964            2019  
3965            2019  
3966            2019  

[3967 rows x 8 columns]
```

[123]:
```
df.dtypes
```

[123]:
```
Utility ID                   int64
Plant Code                   int64
State                       object
Nameplate Capacity (MW)    float64
```

```
Summer Capacity (MW)       float64
Winter Capacity (MW)        object
Operating Month              int64
Operating Year               int64
dtype: object
```

```
[146]: summer = df.columns[4]
       winter = df.columns[5]
```

```
[136]: winter
```

```
[136]: 'Winter Capacity (MW)'
```

```
[145]: #df[[winter]] = df[[winter]].astype('float')
```

```
[142]: df[winter] = pd.to_numeric(df[winter],errors='coerce')
```

```
[144]: df.dtypes
```

```
[144]: Utility ID                   int64
       Plant Code                   int64
       State                       object
       Nameplate Capacity (MW)    float64
       Summer Capacity (MW)       float64
       Winter Capacity (MW)       float64
       Operating Month              int64
       Operating Year               int64
       dtype: object
```

```
[147]: df['Total Capacity'] = df[summer] +df[winter]
```

```
[229]: group = df.groupby('State').agg({'Total Capacity':['sum', 'mean']})
       group
```

```
[229]:       Total Capacity
                        sum         mean
       State
       AL             385.2    64.200000
       AR             224.4    28.050000
       AZ            4309.8    34.478400
       CA           25039.1    30.461192
       CO            1211.6    12.490722
       CT             274.5     6.238636
       DC              13.8     6.900000
       DE              72.5     6.590909
       FL            4284.3    65.912308
       GA            3026.0    44.500000
```

```
HI            535.4   12.168182
IA             25.8    3.685714
ID            472.0   47.200000
IL             87.6    7.963636
IN            475.4    6.889855
KS             20.4    4.080000
KY             52.0    8.666667
LA              2.2    2.200000
MA           1674.2    4.866860
MD            627.5    7.843750
ME             10.0    5.000000
MI            202.8   11.266667
MN           1788.5    4.064773
MO            124.2    6.536842
MS            441.4   63.057143
MT             34.0    5.666667
NC           8881.7   14.279260
NE             44.6    5.575000
NJ           1689.2    5.844983
NM           1319.2   19.118841
NV           4617.0   69.954545
NY            957.3    5.318333
OH            212.4    7.080000
OK             61.0    8.714286
OR            800.2   14.289286
PA            157.4    4.769697
RI            159.0    7.571429
SC           1311.3   21.150000
SD              2.0    2.000000
TN            360.0   20.000000
TX           4893.0   77.666667
UT           1832.5   57.265625
VA           1022.2   51.110000
VT            241.8    6.045000
WA             39.4   19.700000
WI             80.8    3.672727
WY            184.0  184.000000
```

[ ]:

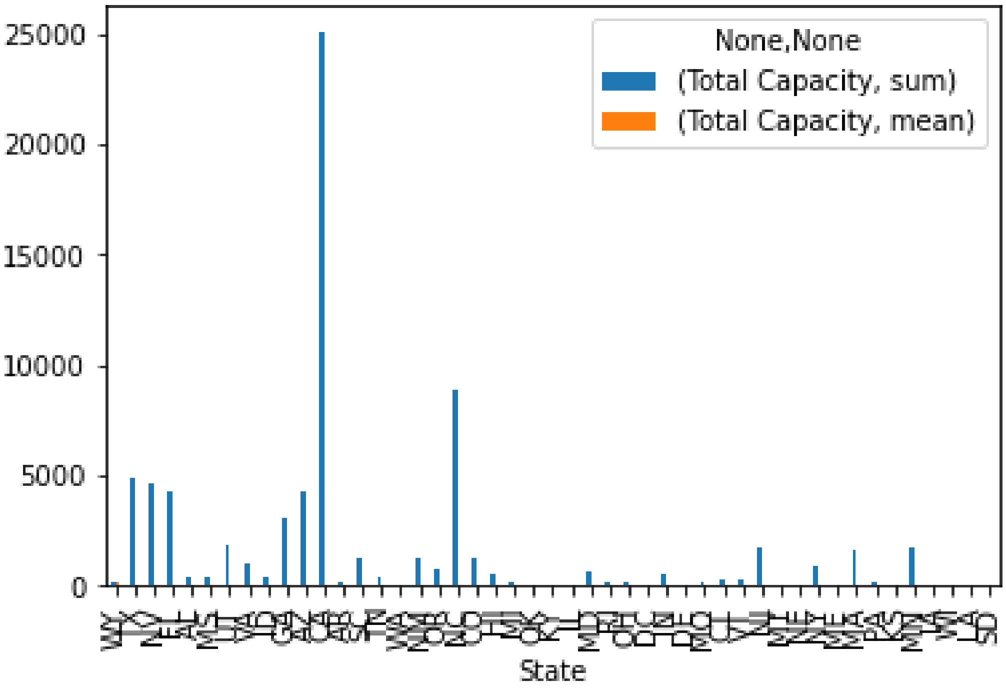
```
```

7 Question 7

[170]:

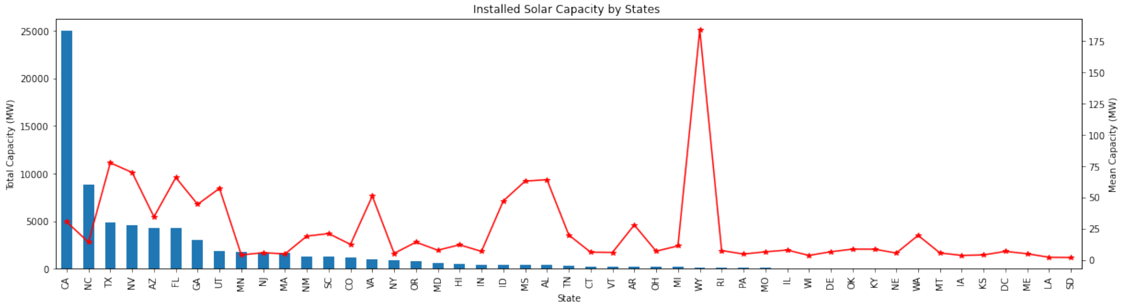
```
ax = group.sort_values(by = [('Total Capacity','mean'), ('Total Capacity','sum')], ascending =False).plot(kind = 'bar')
ax.set_ylabel()
```

```
[170]: <matplotlib.axes._subplots.AxesSubplot at 0x7f548ea35d60>
```



```
[246]: plt.figure(figsize= (20,5))
        ax = group.iloc[:,0].sort_values(ascending =False).plot(kind = 'bar')
        ax.set_ylabel('Total Capacity (MW)')
        ax1 =ax.twinx()
        ax1=  group.sort_values(by = ('Total Capacity','sum'),ascending =False).iloc[:
         ↪,1].plot(kind = 'line', color ='r', marker = '*')
        ax1.set_ylabel('Mean Capacity (MW)')
        plt.title("Installed Solar Capacity by States ")
```

```
[246]: Text(0.5, 1.0, 'Installed Solar Capacity by States ')
```

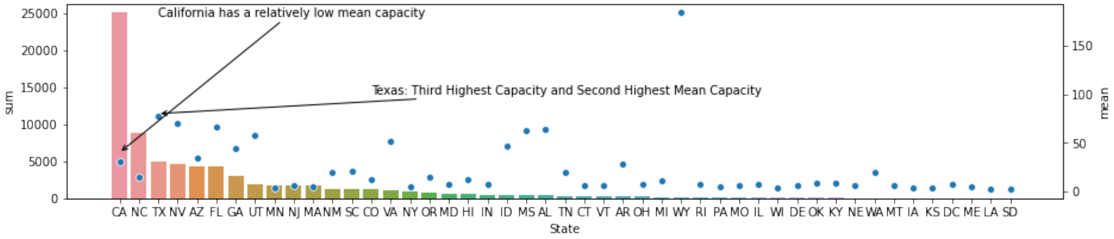


```
[247]: group.columns = group.columns.droplevel(0)
```

```
[248]: group = group.reset_index()
```

```
[285]: plt.figure(figsize = (15,3))
       group = group.sort_values(by='sum', ascending = False)
       ax = sns.barplot(x='State', y='sum', data =group)
       ax1 =ax.twinx()
       ax1 = sns.scatterplot(x= 'State', y ='mean', data =group)
       plt.annotate("Texas: Third Highest Capacity and Second Highest Mean Capacity",␣
        ↪('TX',80), xycoords="data", xytext=('CO',100),␣
        ↪arrowprops=dict(arrowstyle='->'))
       plt.annotate("California has a relatively low mean capacity", ('CA',40),␣
        ↪xycoords="data", xytext=('TX',180), arrowprops=dict(arrowstyle='->'))
```

```
[285]: Text(TX, 180, 'California has a relatively low mean capacity')
```



```
[ ]:
```

```
[ ]:
```

```
[ ]:
```

```
[ ]:
```

```
[ ]:
```

```
[ ]:
```