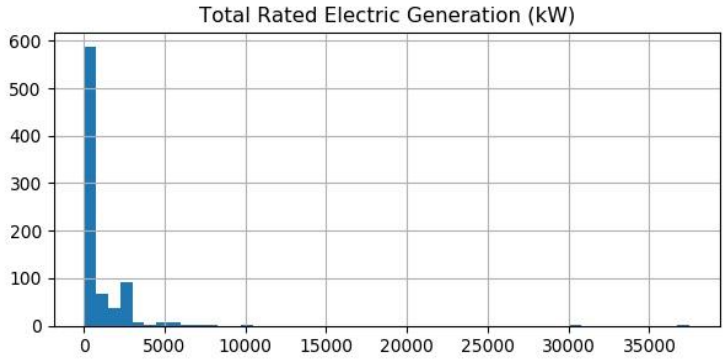
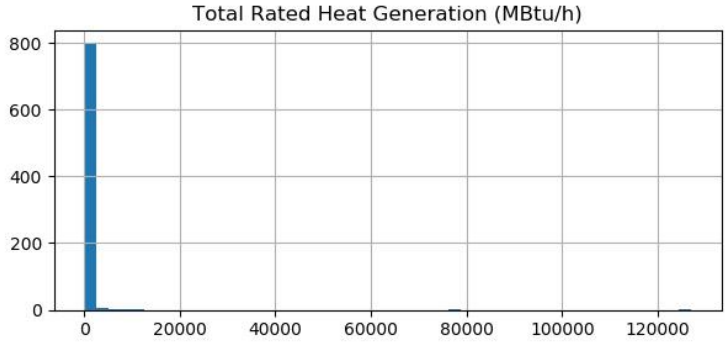
EDA report on
Facilities information for distributed power generation for the state of New York, USA

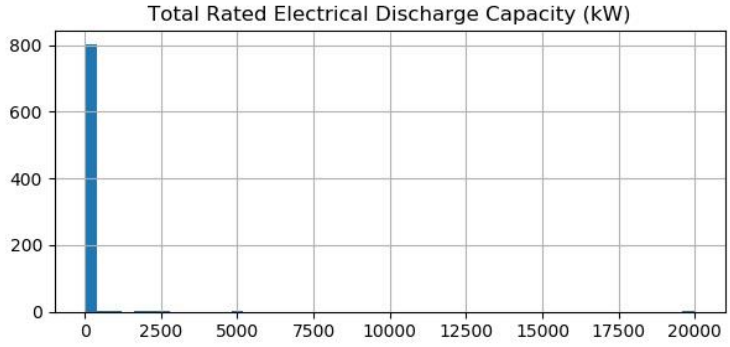
Numerical Distribution:

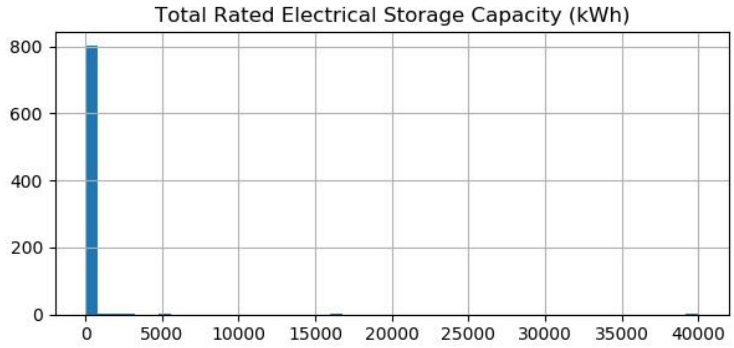
	Count	Mean	Standard deviation	Minimum value	First quartile	Median	Third quartile	Maximum value
Total Rated Electric Generation (kW)	811	876.54	2063.77	0	70	300	912.35	37500
Total Rated Heat Generation (MBtu/h)	811	428.62	5260.08	0	0	0	0	127000
Total Rated Electrical Discharge Capacity (kW)	811	48.15	740.21	0	0	0	0	20000
Total Rated Electrical Storage Capacity (kWh)	811	93.90	1531.45	0	0	0	0	40000
Total Rated Cooling Energy Discharge Capacity (ton)	811	0	0	0	0	0	0	0
Total Equivalent Electrical Discharge Capacity from Cooling Sources (kW)	811	18.38	211.84	0	0	0	0	5000
Total Rated Cooling Energy Storage Capacity (ton-hour)	811	143.55	1433.68	0	0	0	0	30000
Total Equivalent Electrical Storage Capacity from Cooling Sources (kWh)	811	1590.92	43150.59	0	0	0	0	1228680
Total Rated Thermal Energy Discharge Capacity (MBtu/h)	811	0	0	0	0	0	0	0
Total Equivalent Electrical Discharge Capacity from Thermal Sources (kW)	811	0	0	0	0	0	0	0
Total Rated Thermal Energy Storage Capacity (MBtu)	811	0	0	0	0	0	0	0
Total Equivalent Electrical Storage Capacity from Thermal Sources (kWh)	811	0	0	0	0	0	0	0
Installed Systems	811	1.06	0.62	0	1	1	1	11

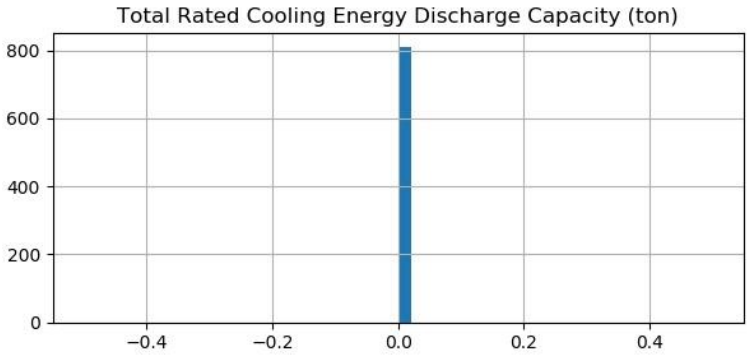
Univariate analysis:

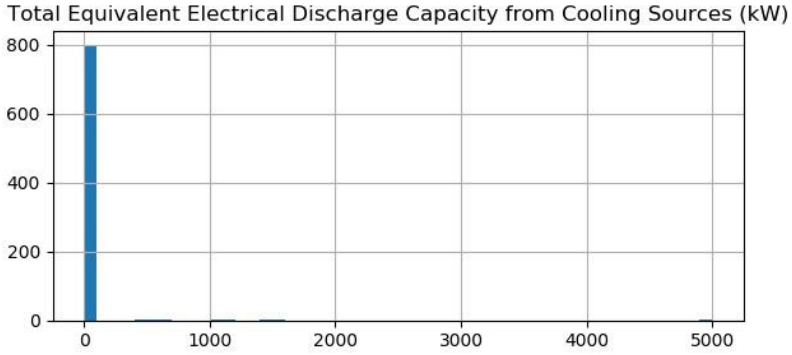


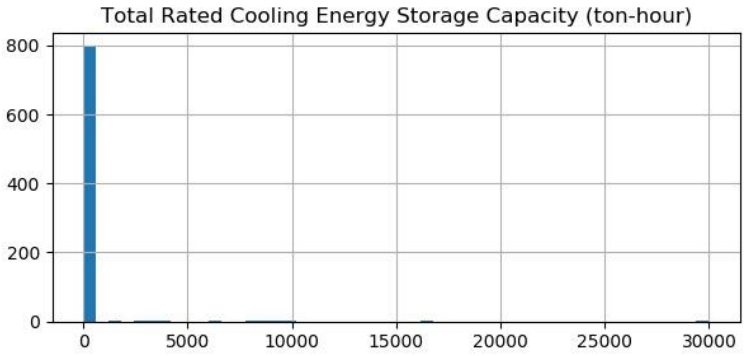


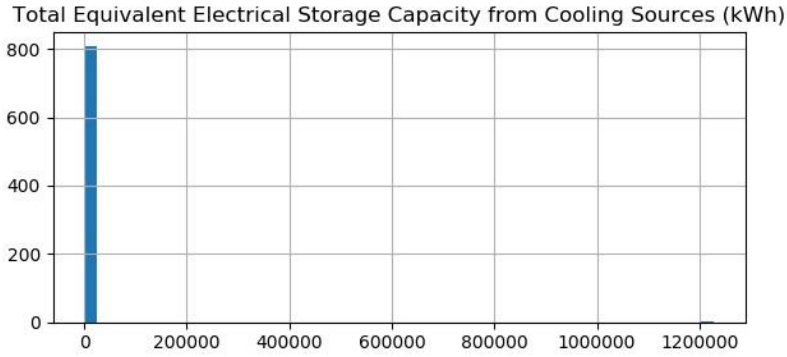


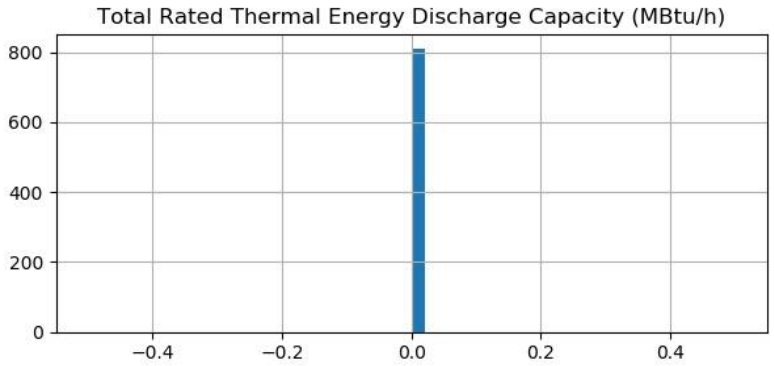


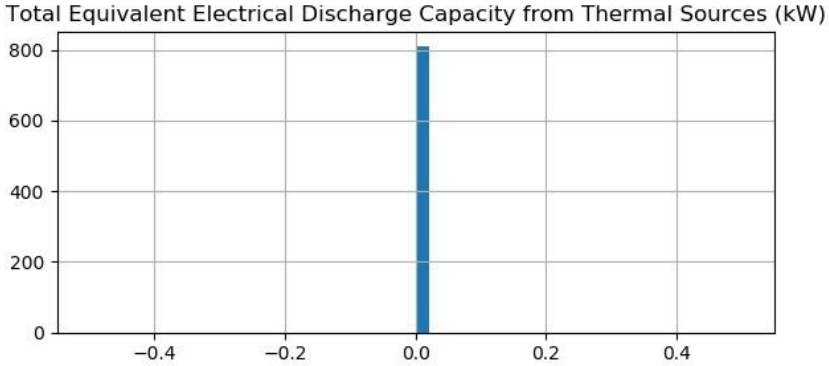

Total Rated Cooling Energy Discharge Capacity (ton)
800
600
400
200
0
−0.4
−0.2
0.0
0.2
0.4

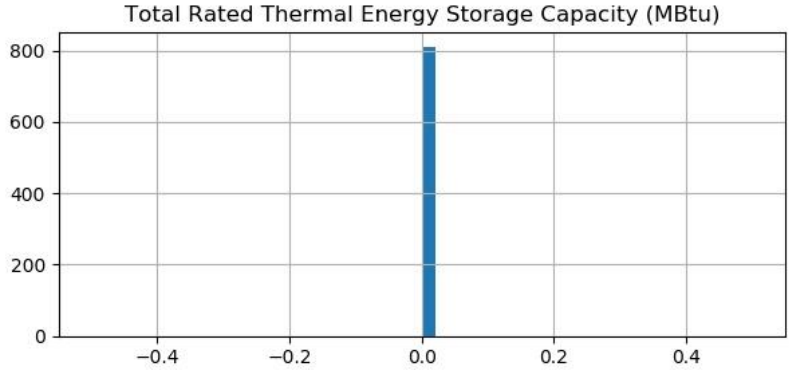

Total Equivalent Electrical Discharge Capacity from Cooling Sources (kW)
800
600
400
200
0
0
1000
2000
3000
4000
5000

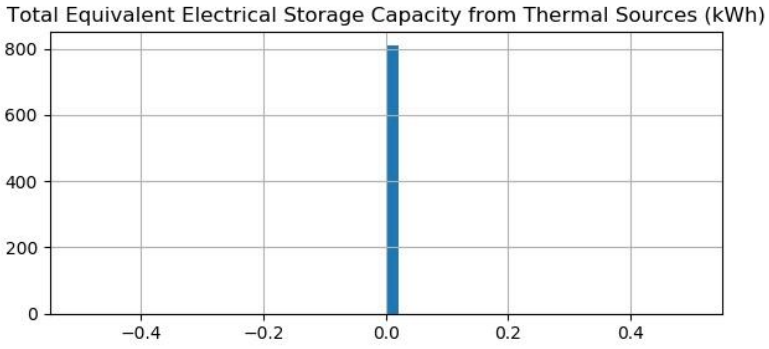

Total Rated Cooling Energy Storage Capacity (ton-hour)
800
600
400
200
0
0
5000
10000
15000
20000
25000
30000

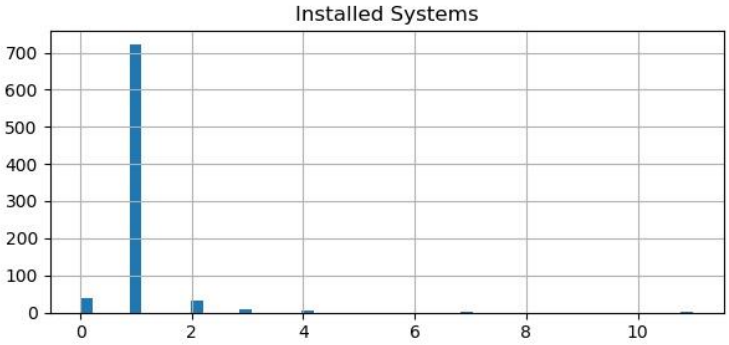

Total Equivalent Electrical Storage Capacity from Cooling Sources (kWh)
800
600
400
200
0
0
200000
400000
600000
800000
1000000
1200000


Total Rated Thermal Energy Discharge Capacity (MBtu/h)
800
600
400
200
0
−0.4
−0.2
0.0
0.2
0.4

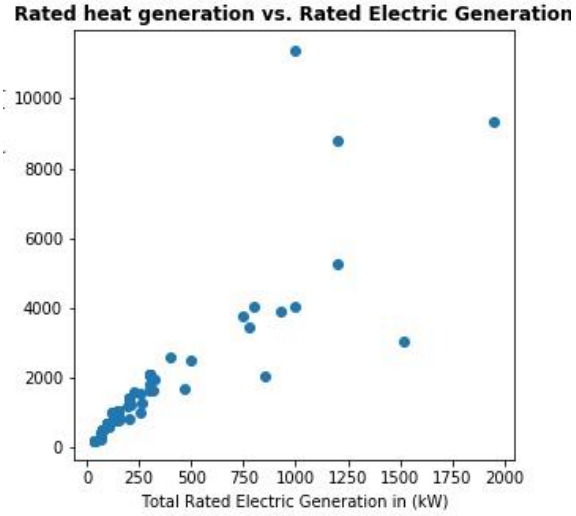

Total Equivalent Electrical Discharge Capacity from Thermal Sources (kW)
800
600
400
200
0
−0.4
−0.2
0.0
0.2
0.4

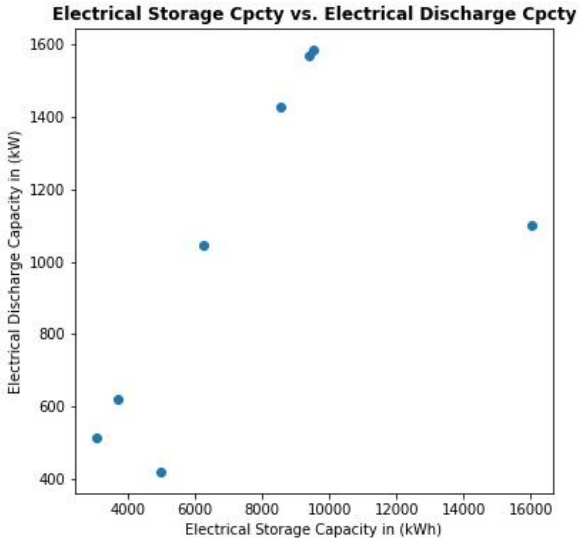

Total Rated Thermal Energy Storage Capacity (MBtu)
800
600
400
200
0
−0.4
−0.2
0.0
0.2
0.4

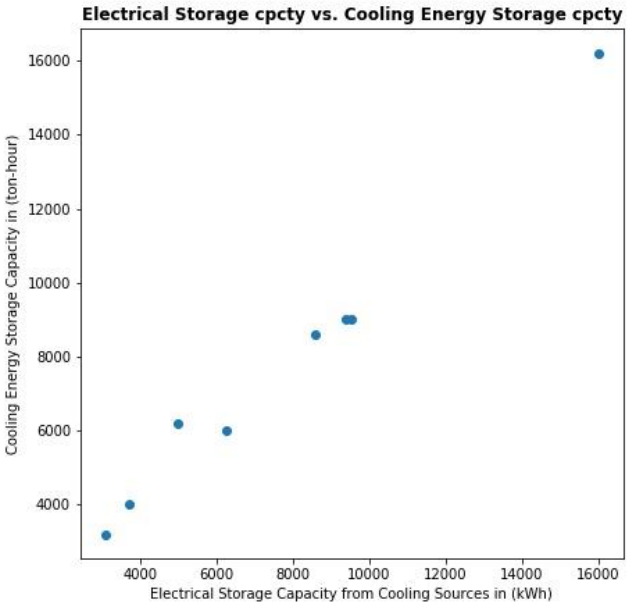

Total Equivalent Electrical Storage Capacity from Thermal Sources (kWh)
800
600
400
200
0
−0.4
−0.2
0.0
0.2
0.4


Installed Systems
700
600
500
400
300
200
100
0
0
2
4
6
8
10

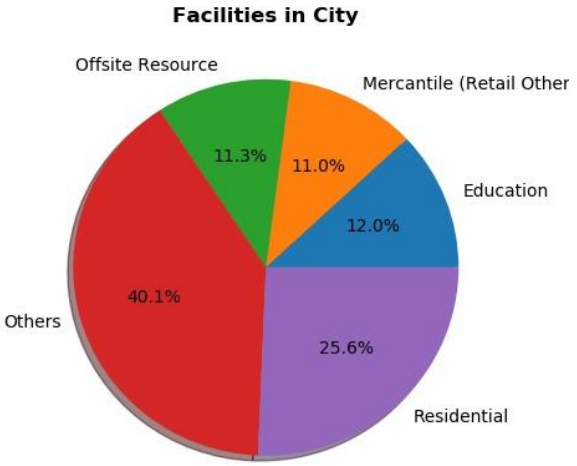
Bivariate analysis

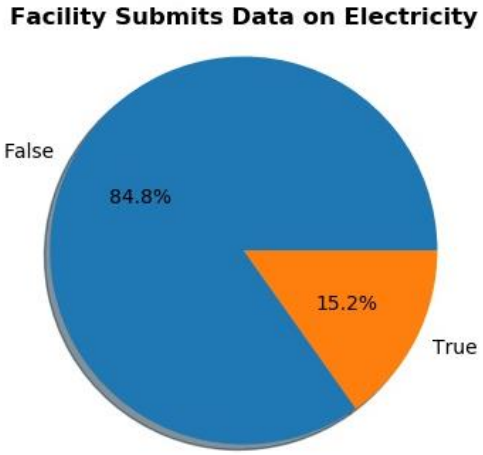


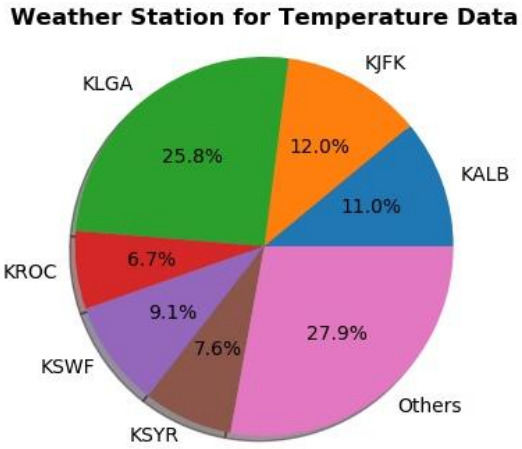


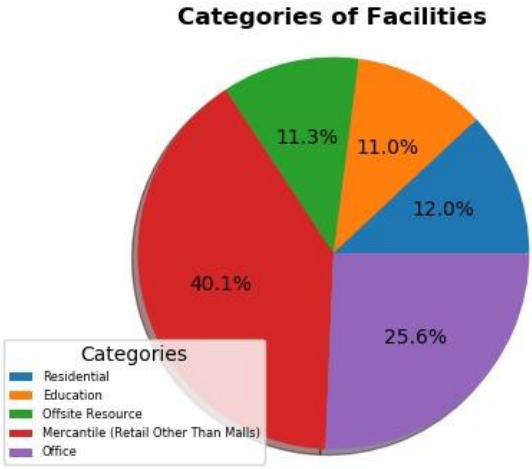


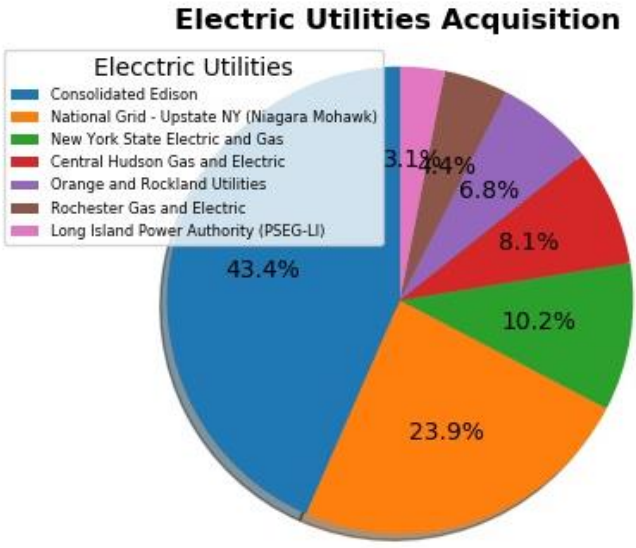
Categorical Value analysis:











Insights into dataset:

- The dataset contains 811 observations and 35 variables
- Most facilities have only one line Address
- Most facilities do not have Alternate Name(s)
- Most facilities do not have Floor Area record.
- Number of Occupancy Units are not mentioned for most of the facilities.
- No facilities have data for Floor Plan
- The Latitude Longitude coordinates confirm that most of the facilities are located in and around New York.
- The following parameters are not rated for any Facilities:
    - Cooling energy discharge capacity
    - Thermal energy storage capacity

Executive Summary:

From the correlation matrix, it is clear that there is strong relation between Electric generation and Heat generation. Thus, increase in electric generation would mean increase in heat generation. Hence, a robust cooling system is required, i.e., a system with more Rated Cooling Energy Storage Capacity.

Also, to increase Electric Storage Capacity, we need to have system with more Electric Discharge Capacity, as seen from the scatter plot.

Tools Used:

- Jupyter Notebook
- Python 3.6.3
- pandas 0.20.3
- numpy 1.13.3
- matplotlib 2.1.0

Methodology used:

1. Import required package
2. Fetch the 'Facilities' data from the given excel file and store it in a pandas dataframe.
3. Get an overview of data, like no. of columns and rows, no. of missing values, type of data.
4. Remove columns/rows which are not required.
5. Examine the numerical variables.
6. Generated histograms of relevant variables, so that we can get an idea of skewness.
7. Generated correlation matrix.
8. Generated scatter plot to visualize the correlations. We needed to remove outliers and 0 values to get a clear picture.
9. Examine the categorical values.
10. Generated pie charts to show the contribution of each categorical variables.