Capstone Project 1: Data Visualization

Appliances Energy Prediction –

Business Problem Description – Dataset contains the house temperature and humidity conditions were monitored with a ZigBee wireless sensor network. As per the description on UCI website, each wireless node transmitted the temperature and humidity conditions around 3.3 min, then, the wireless data was averaged for 10 minutes periods. The energy data was logged every 10 minutes with m-bus energy meters. Combining this data with the weather data based on the date time columns

Link for the Jupyter file
- https://github.com/arijitsinha80/Springboard/blob/master/Project/Capstoneproject_Phase2.ipynb

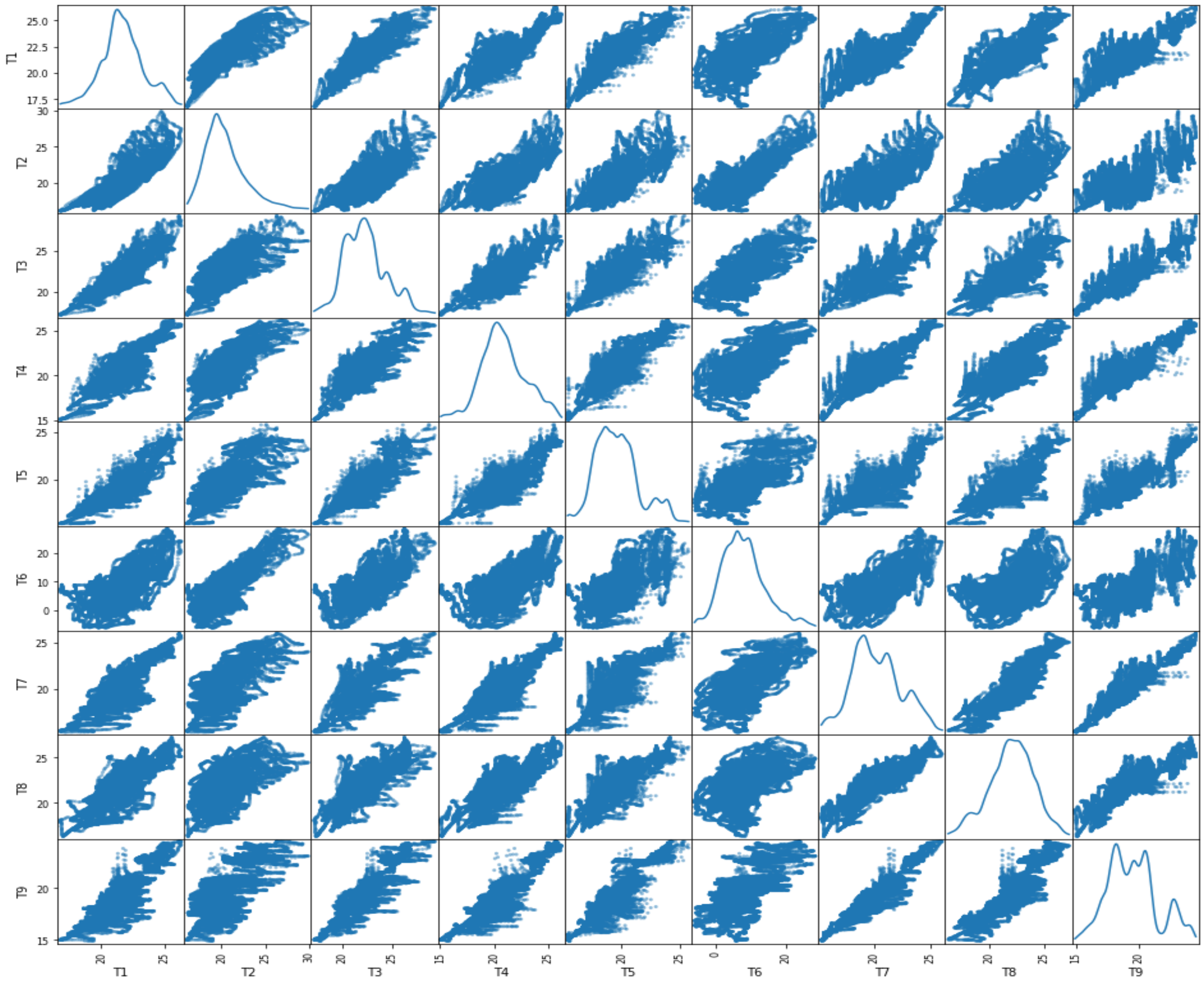
From the Data Wrangling activity, we created the **input.csv** as the final dataset. This has 19735 observations and 30 attributes.

Divide the data in dimension wise to explore from the input dataset -

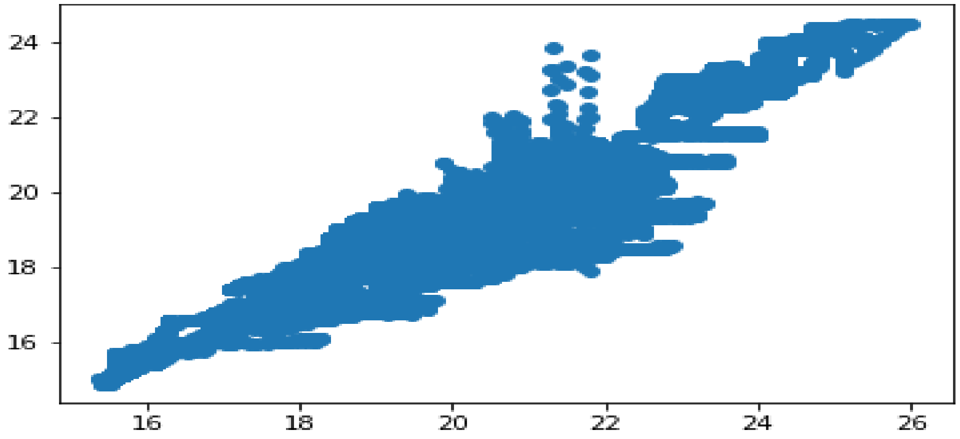
```
# Temperature sensors columns
temp_cols = ["T1", "T2", "T3", "T4", "T5", "T6", "T7", "T8", "T9"]

# Humidity sensors columns
hum_cols = ["RH_1", "RH_2", "RH_3", "RH_4", "RH_5", "RH_6", "RH_7", "RH_8"
, "RH_9"]

# Weather data columns
wth_cols = ["T_out", "Tdewpoint", "RH_out", "Press_mm_hg", "Windspeed", "V
isibility"]

# Target column
tgt = ["Appliances"]
```

From the above dimensions, we will start to explore data for each –

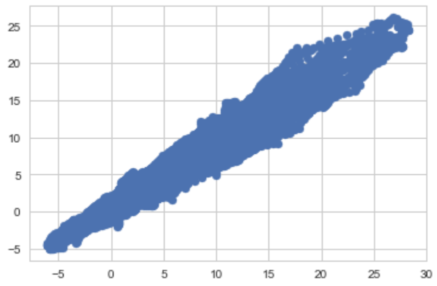
Plot the scatter matrix for Temperature attributes using method “ **diagonal="kde"**



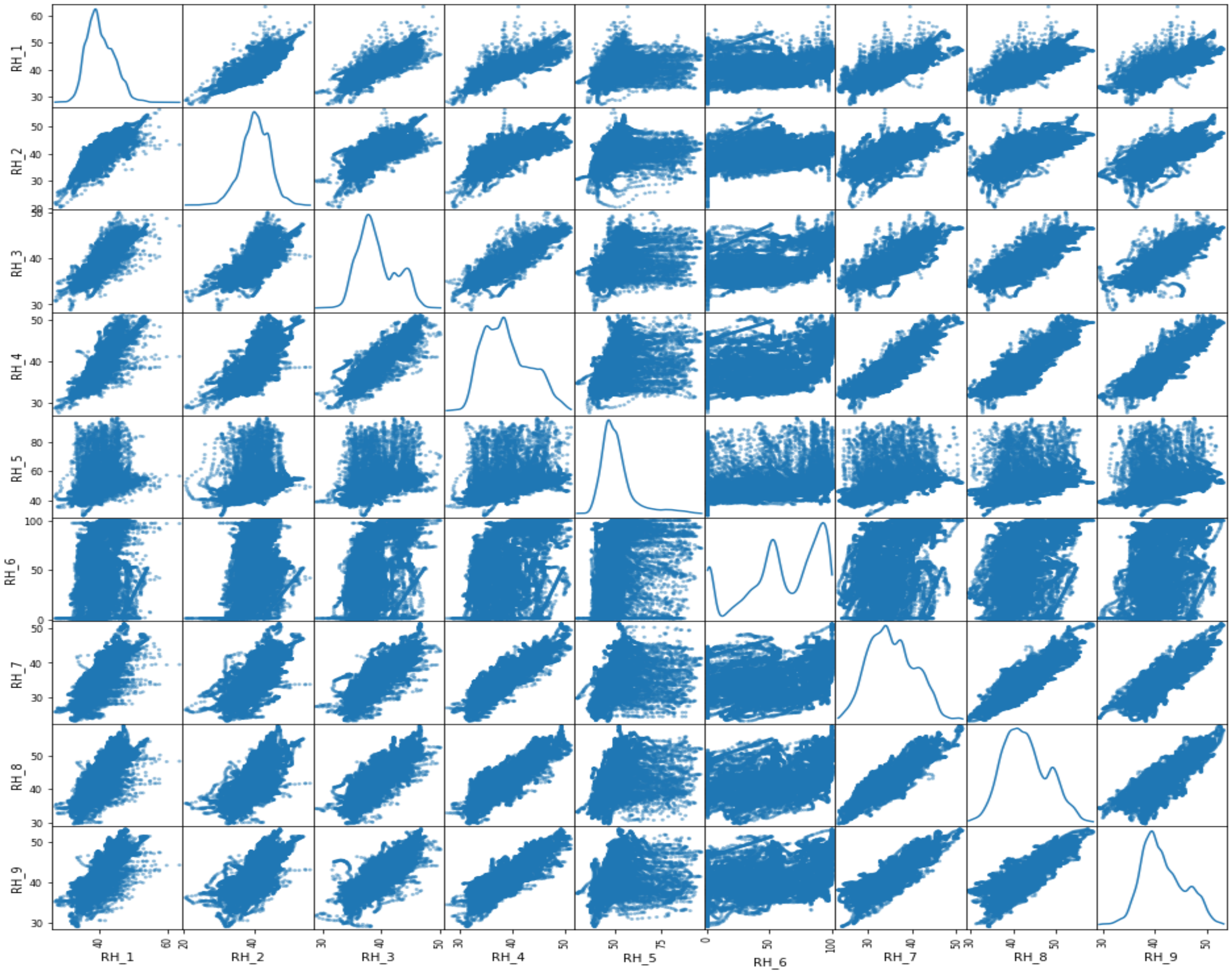
- From the above figure, we can see that there is some linear relation between T7 and T9. Others are having the shape but are not exactly linear.
- There is a relation between these two attributes but also have some outliers



T6 and T_out is highly correlated, T6 is from the outside the house reading and T_out is the data collected from weather's site

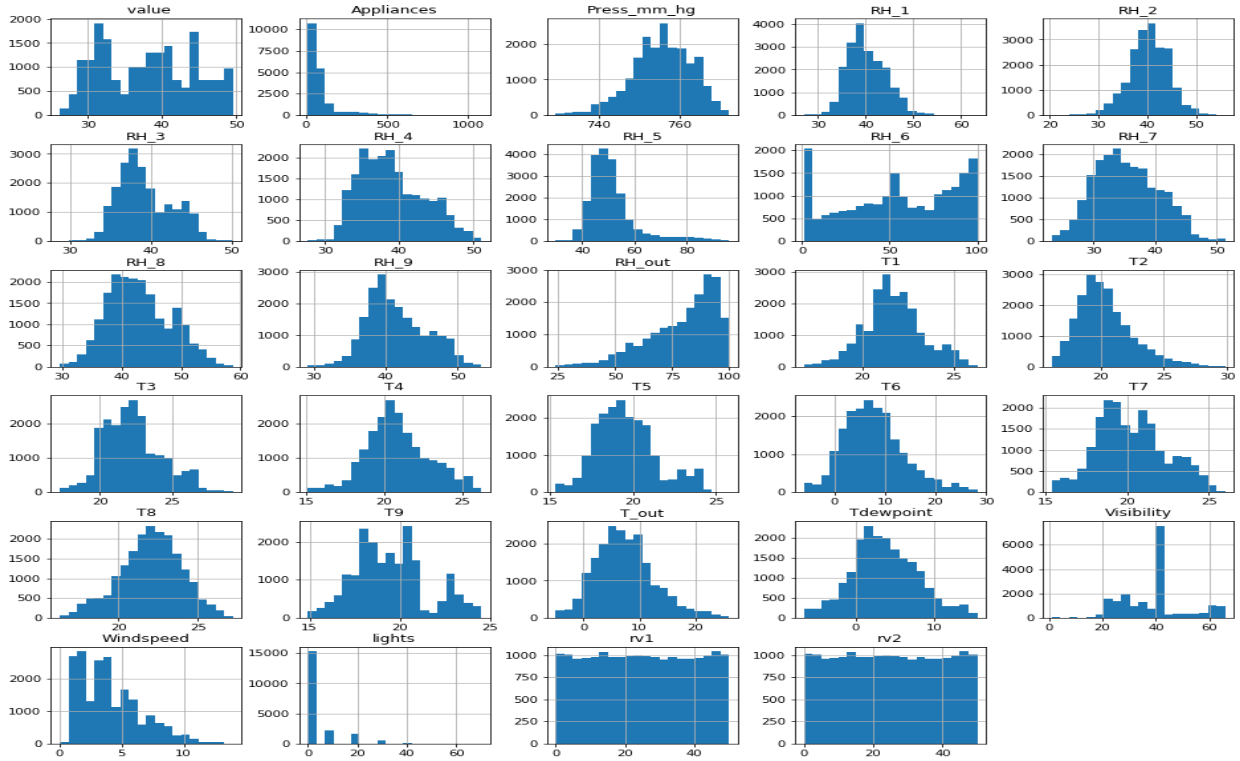


Explore the data using the Weather Dimension -



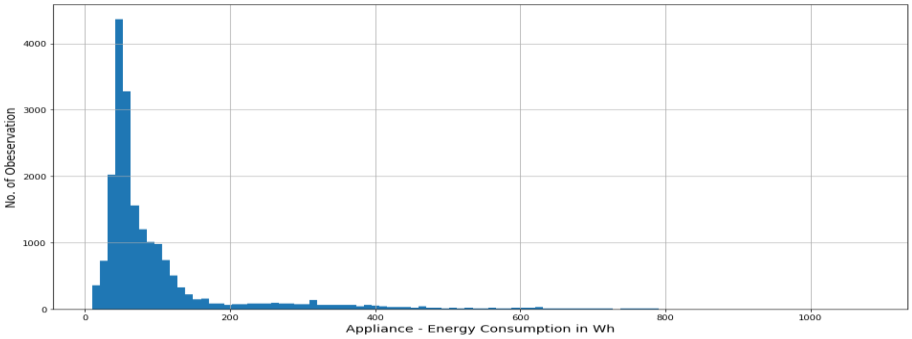
- There doesn’t seems to be having any linearity between any of the attributes.

Lets explore the distribution using the histogram –

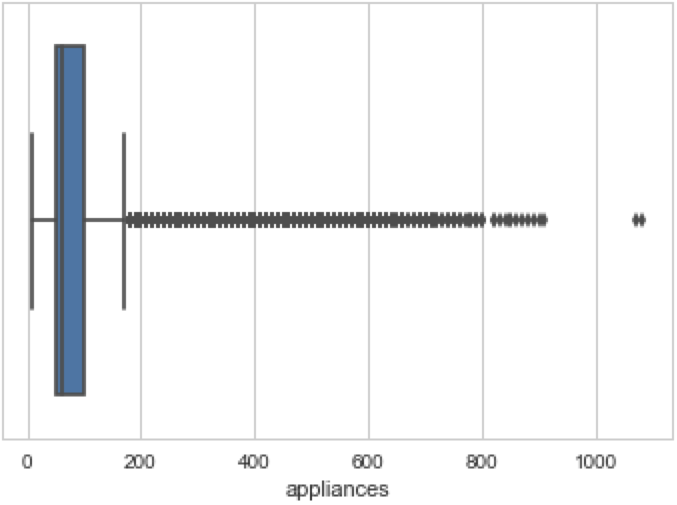


- All humidity values are almost having normal distribution except RH_6 and RH_out. In other words the reading from inside the home is having normal distribution.
- All temperature readings follow a Normal distribution except for T9.
- Visibility, Windspeed and Appliances are having skewed data.
- Rv1 and Rv2 are random variables and doesn't seems to be contributing

On the Target Attribute – Appliance, the below histograms is rightly skewed and most of the data is with 200 KWh.



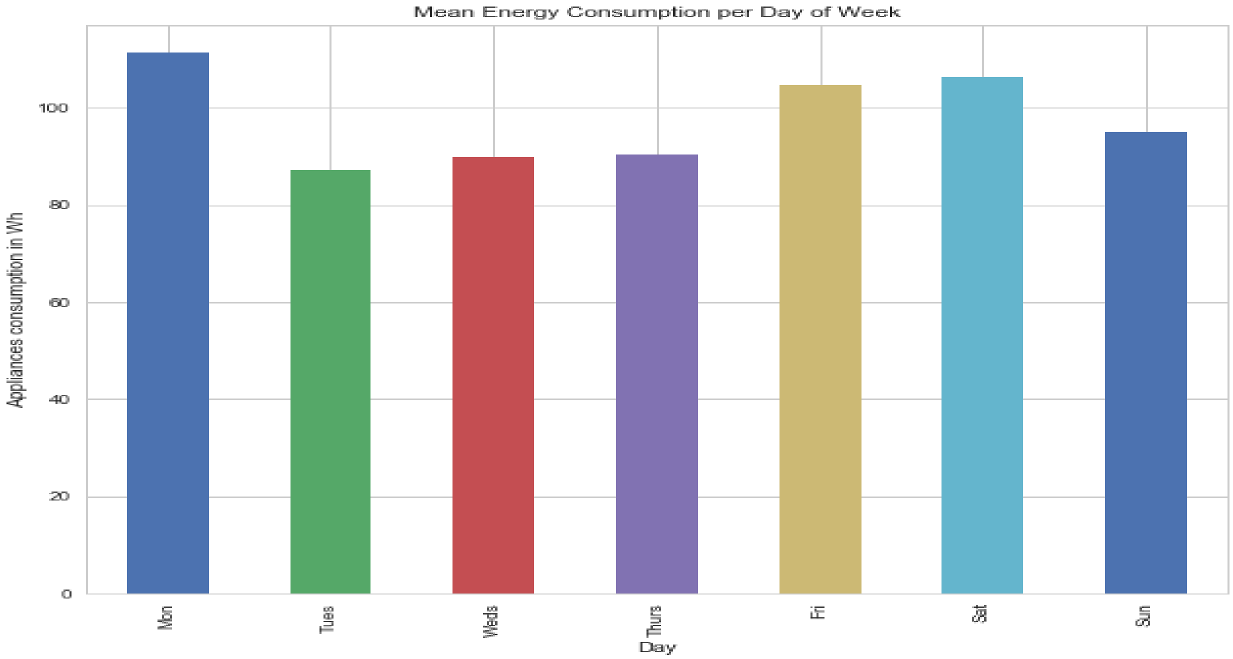
Target variable, Appliances is highly right skewed.
Alternatively exploring using Boxplot – on Appliance Attribute



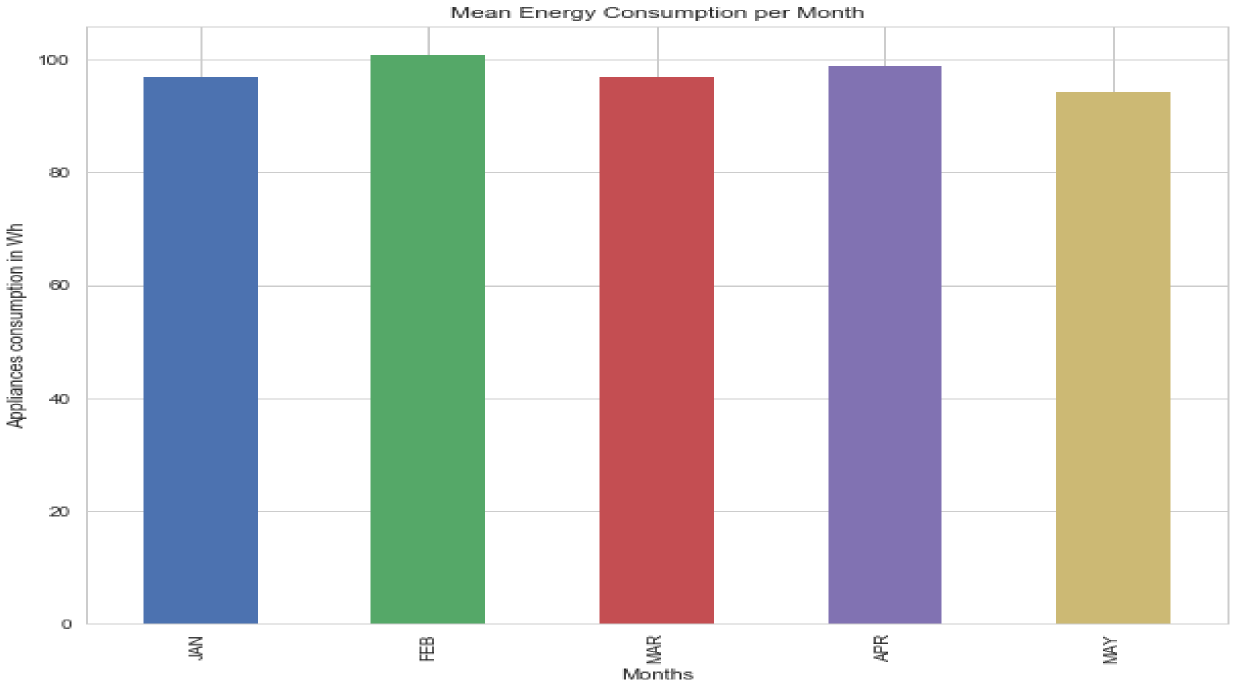
- Percentage of dataset in range of 0-200 KWh is 90.291%

Using the date attributes, created new columns for Month and Weeks using the **datetimestamp** method

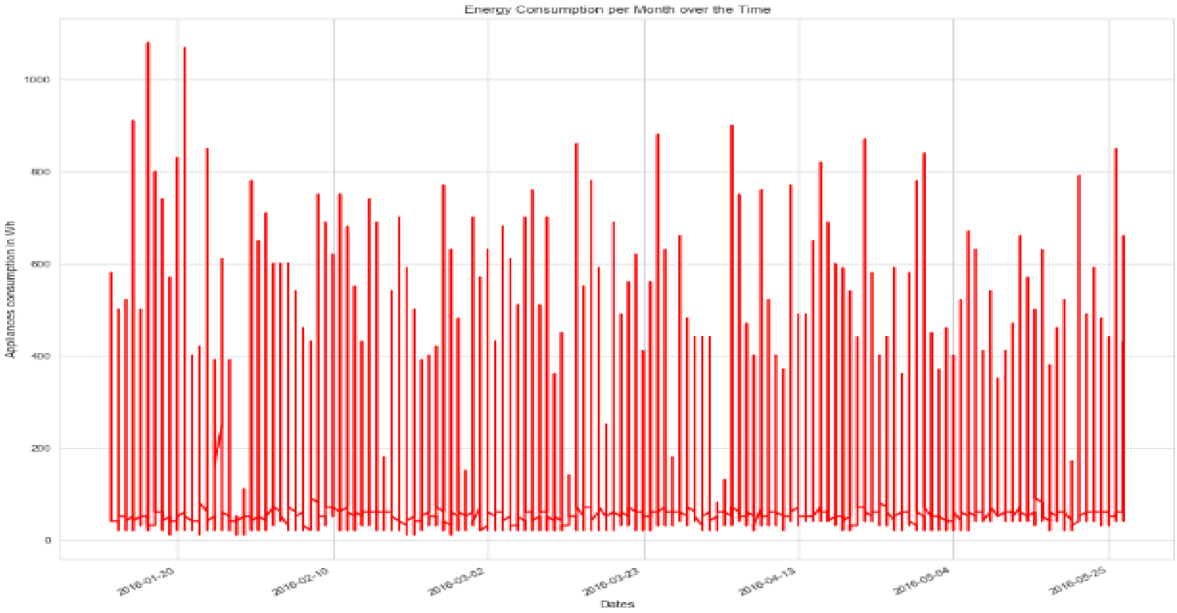
With taking the average on week – Monday the usage has been higher, followed by Saturday and Friday.



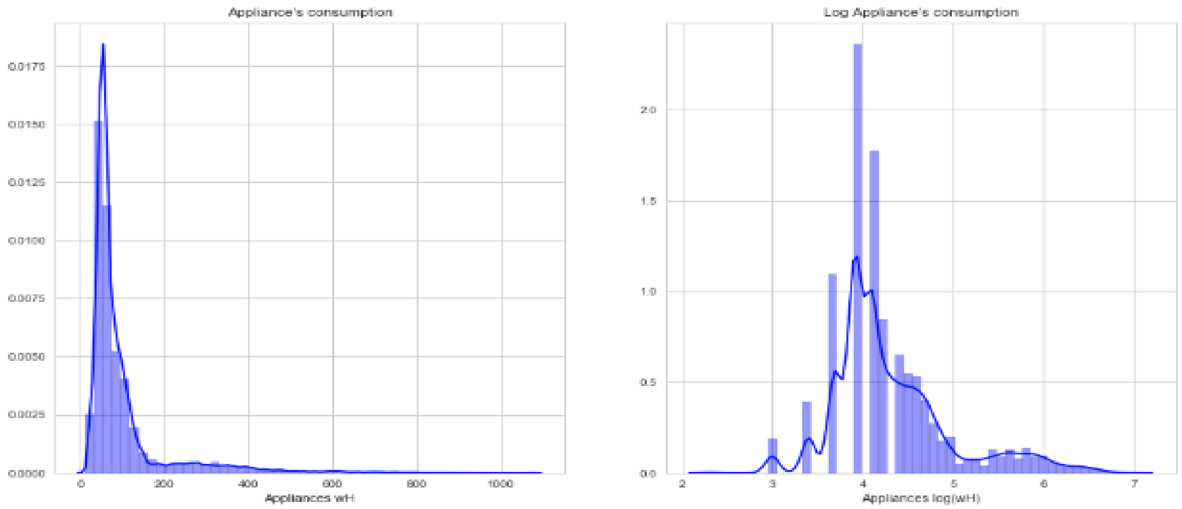
For Monthly Average – On a average, February and April the consumption has been more than other months.



For Day wise consumption – plotting this date wise, energy consumption, In January month there were 2 days when the consumption was more than 1000 KWh.

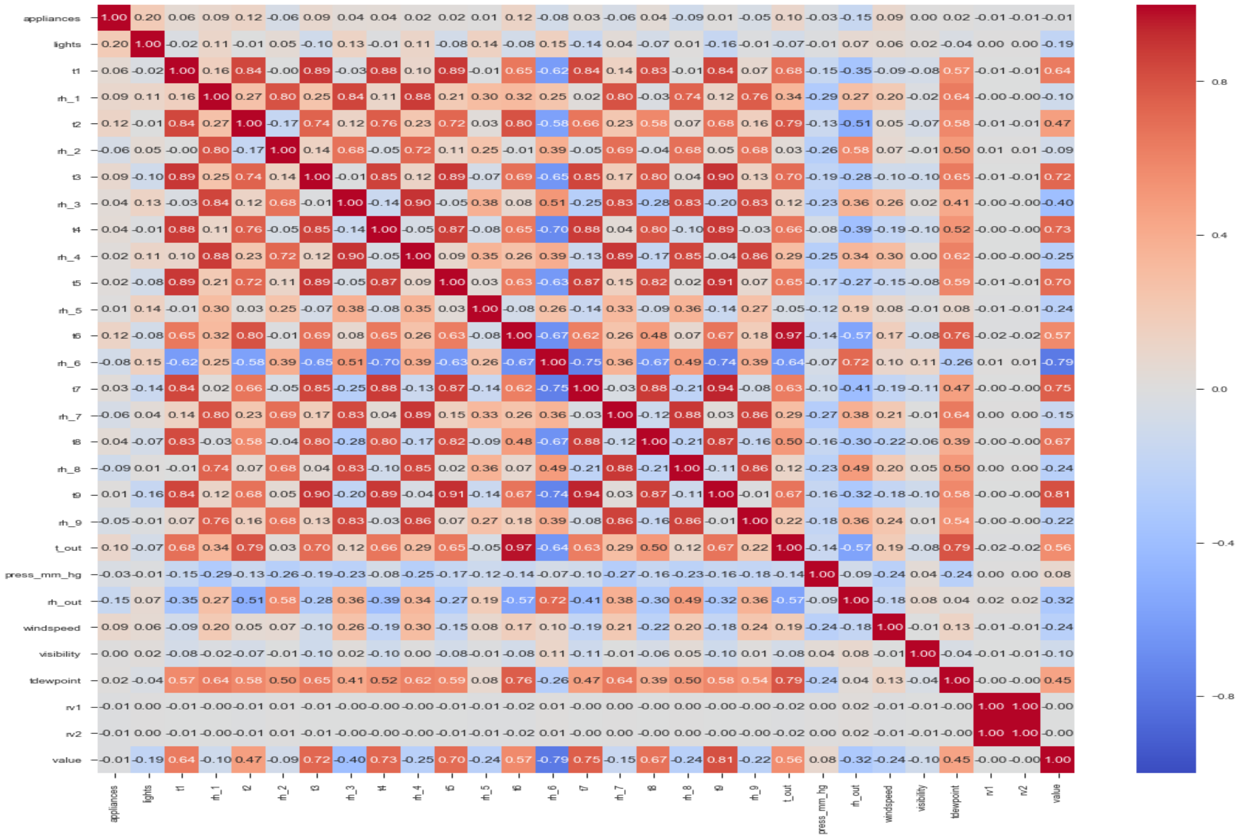


Since the Appliance data captured were rightly skewed, converting the column to Log values to see if it has the normal distribution.



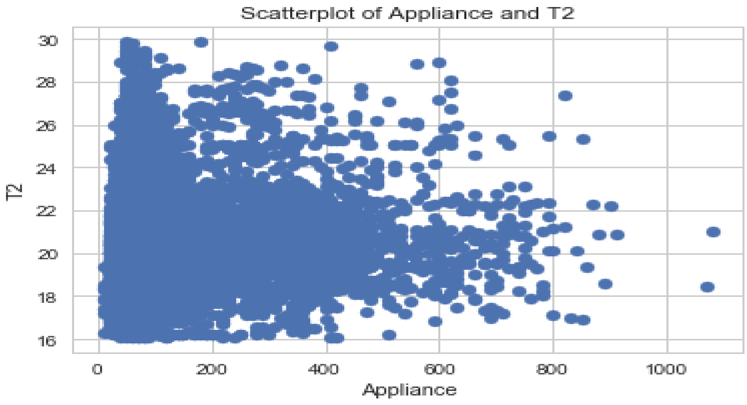
Let's explore the Correlation plot –

With Appliance attribute –

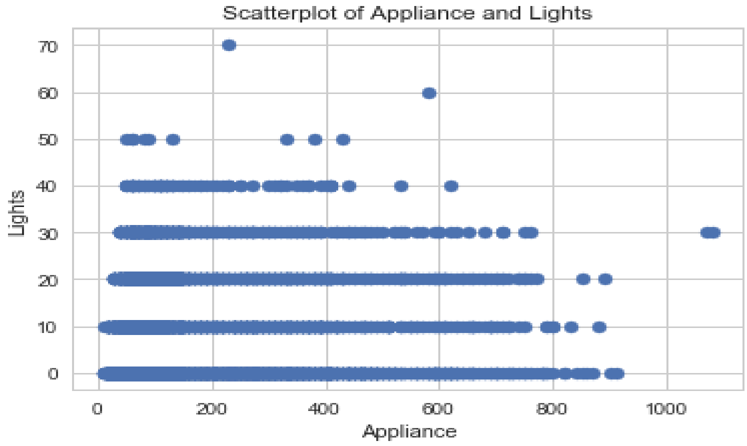


	appliances	lights	t1	rh_1	t2	rh_2	t3	rh_3	t4	rh_4	t5	rh_5	t6	rh_6	t7	rh_7	t8	rh_8	t9	rh_9	t_out	press_mm_hg	rh_out	windspeed	visibility	tdewpoint	rv1	rv2	value
appliances	1.00	0.20	0.06	0.09	0.12	-0.06	0.09	0.04	0.04	0.02	0.02	0.01	0.12	-0.08	0.03	-0.06	0.04	-0.09	0.01	-0.05	0.10	-0.03	-0.15	0.09	0.00	0.02	-0.01	-0.01	-0.01
lights	0.20	1.00	-0.02	0.11	-0.01	0.05	-0.10	0.13	-0.01	0.11	-0.08	0.14	-0.08	0.15	-0.14	0.04	-0.07	0.01	-0.16	-0.01	-0.07	-0.01	0.07	0.06	0.02	-0.04	0.00	0.00	-0.19
t1	0.06	-0.02	1.00	0.16	0.84	-0.00	0.89	-0.03	0.88	0.10	0.89	-0.01	0.65	-0.62	0.84	0.14	0.83	-0.01	0.84	0.07	0.68	-0.15	-0.35	-0.09	-0.08	0.57	-0.01	-0.01	0.64
rh_1	0.09	0.11	0.16	1.00	0.27	0.80	0.25	0.84	0.11	0.88	0.21	0.30	0.32	0.25	0.02	0.80	-0.03	0.74	0.12	0.76	0.34	-0.29	0.27	0.20	-0.02	0.64	-0.00	-0.00	-0.10
t2	0.12	-0.01	0.84	0.27	1.00	-0.17	0.74	0.12	0.76	0.23	0.72	0.03	0.80	-0.58	0.66	0.23	0.58	0.07	0.68	0.16	0.79	-0.13	-0.51	0.05	-0.07	0.58	-0.01	-0.01	0.47
rh_2	-0.06	0.05	-0.00	0.80	-0.17	1.00	0.14	0.68	-0.05	0.72	0.11	0.25	-0.01	0.39	-0.05	0.69	-0.04	0.68	0.05	0.68	0.03	-0.26	0.58	0.07	-0.01	0.50	0.01	0.01	-0.09
t3	0.09	-0.10	0.89	0.25	0.74	0.14	1.00	-0.01	0.85	0.12	0.89	-0.07	0.69	-0.65	0.85	0.17	0.80	0.04	0.90	0.13	0.70	-0.19	-0.28	-0.10	-0.10	0.65	-0.01	-0.01	0.72
rh_3	0.04	0.13	-0.03	0.84	0.12	0.68	-0.01	1.00	-0.14	0.90	-0.05	0.38	0.08	0.51	-0.25	0.83	-0.28	0.83	-0.20	0.83	0.12	-0.23	0.36	0.26	0.02	0.41	-0.00	-0.00	-0.40
t4	0.04	-0.01	0.88	0.11	0.76	-0.05	0.85	-0.14	1.00	-0.05	0.87	-0.08	0.65	-0.70	0.88	0.04	0.80	-0.10	0.89	-0.03	0.66	-0.08	-0.39	-0.19	-0.10	0.52	-0.00	-0.00	0.73
rh_4	0.02	0.11	0.10	0.88	0.23	0.72	0.12	0.90	-0.05	1.00	0.09	0.35	0.26	0.39	-0.13	0.89	-0.17	0.85	-0.04	0.86	0.29	-0.25	0.34	0.30	0.00	0.62	-0.00	-0.00	-0.25
t5	0.02	-0.08	0.89	0.21	0.72	0.11	0.89	-0.05	0.87	0.09	1.00	0.03	0.63	-0.63	0.87	0.15	0.82	0.02	0.91	0.07	0.65	-0.17	-0.27	-0.15	-0.08	0.59	-0.01	-0.01	0.70
rh_5	0.01	0.14	-0.01	0.30	0.03	0.25	-0.07	0.38	-0.08	0.35	0.03	1.00	-0.08	0.26	-0.14	0.33	-0.09	0.36	-0.14	0.27	-0.05	-0.12	0.19	0.08	-0.01	0.08	-0.01	-0.01	-0.24
t6	0.12	-0.08	0.65	0.32	0.80	-0.01	0.69	0.08	0.65	0.26	0.63	-0.08	1.00	-0.67	0.62	0.26	0.48	0.07	0.67	0.18	0.97	-0.14	-0.57	0.17	-0.08	0.76	-0.02	-0.02	0.57
rh_6	-0.08	0.15	-0.62	0.25	-0.58	0.39	-0.65	0.51	-0.70	0.39	-0.63	0.26	-0.67	1.00	-0.75	0.36	-0.67	0.49	-0.74	0.39	-0.64	-0.07	0.72	0.10	0.11	-0.26	0.01	0.01	-0.79
t7	0.03	-0.14	0.84	0.02	0.66	-0.05	0.85	-0.25	0.88	-0.13	0.87	-0.14	0.62	-0.75	1.00	-0.03	0.88	-0.21	0.94	-0.08	0.63	-0.10	-0.41	-0.19	-0.11	0.47	-0.00	-0.00	0.75
rh_7	-0.06	0.04	0.14	0.80	0.23	0.69	0.17	0.83	0.04	0.89	0.15	0.33	0.26	0.36	-0.03	1.00	-0.12	0.88	0.03	0.86	0.29	-0.27	0.38	0.21	-0.01	0.64	0.00	0.00	-0.15
t8	0.04	-0.07	0.83	-0.03	0.58	-0.04	0.80	-0.28	0.80	-0.17	0.82	-0.09	0.48	-0.67	0.88	-0.12	1.00	-0.21	0.87	-0.16	0.50	-0.16	-0.30	-0.22	-0.06	0.39	-0.00	-0.00	0.67
rh_8	-0.09	0.01	-0.01	0.74	0.07	0.68	0.04	0.83	-0.10	0.85	0.02	0.36	0.07	0.49	-0.21	0.88	-0.21	1.00	-0.11	0.86	0.12	-0.23	0.49	0.20	0.05	0.50	0.00	0.00	-0.24
t9	0.01	-0.16	0.84	0.12	0.68	0.05	0.90	-0.20	0.89	-0.04	0.91	-0.14	0.67	-0.74	0.94	0.03	0.87	-0.11	1.00	-0.01	0.67	-0.16	-0.32	-0.18	-0.10	0.58	-0.00	-0.00	0.81
rh_9	-0.05	-0.01	0.07	0.76	0.16	0.68	0.13	0.83	-0.03	0.86	0.07	0.27	0.18	0.39	-0.08	0.86	-0.16	0.86	-0.01	1.00	0.22	-0.18	0.36	0.24	0.01	0.54	-0.00	-0.00	-0.22
t_out	0.10	-0.07	0.68	0.34	0.79	0.03	0.70	0.12	0.66	0.29	0.65	-0.05	0.97	-0.64	0.63	0.29	0.50	0.12	0.67	0.22	1.00	-0.14	-0.57	0.19	-0.08	0.79	-0.02	-0.02	0.56
press_mm_hg	-0.03	-0.01	-0.15	-0.29	-0.13	-0.26	-0.19	-0.23	-0.08	-0.25	-0.17	-0.12	-0.14	-0.07	-0.10	-0.27	-0.16	-0.23	-0.16	-0.18	-0.14	1.00	-0.09	-0.24	0.04	-0.24	0.00	0.00	0.08
rh_out	-0.15	0.07	-0.35	0.27	-0.51	0.58	-0.28	0.36	-0.39	0.34	-0.27	0.19	-0.57	0.72	-0.41	0.38	-0.30	0.49	-0.32	0.36	-0.57	-0.09	1.00	-0.18	0.08	0.04	0.02	0.02	-0.32
windspeed	0.09	0.06	-0.09	0.20	0.05	0.07	-0.10	0.26	-0.19	0.30	-0.15	0.08	0.17	0.10	-0.19	0.21	-0.22	0.20	-0.18	0.24	0.19	-0.24	-0.18	1.00	-0.01	0.13	-0.01	-0.01	-0.24
visibility	0.00	0.02	-0.08	-0.02	-0.07	-0.01	-0.10	0.02	-0.10	0.00	-0.08	-0.01	-0.08	0.11	-0.11	-0.01	-0.06	0.05	-0.10	0.01	-0.08	0.04	0.08	-0.01	1.00	-0.04	-0.01	-0.01	-0.10
tdewpoint	0.02	-0.04	0.57	0.64	0.58	0.50	0.65	0.41	0.52	0.62	0.59	0.08	0.76	-0.26	0.47	0.64	0.39	0.50	0.58	0.54	0.79	-0.24	0.04	0.13	-0.04	1.00	-0.00	-0.00	0.45
rv1	-0.01	0.00	-0.01	-0.00	-0.01	0.01	-0.01	-0.00	-0.00	-0.00	-0.01	-0.01	-0.02	0.01	-0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.02	0.00	0.02	-0.01	-0.01	-0.00	1.00	1.00	-0.00
rv2	-0.01	0.00	-0.01	-0.00	-0.01	0.01	-0.01	-0.00	-0.00	-0.00	-0.01	-0.01	-0.02	0.01	-0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.02	0.00	0.02	-0.01	-0.01	-0.00	1.00	1.00	-0.00
value	-0.01	-0.19	0.64	-0.10	0.47	-0.09	0.72	-0.40	0.73	-0.25	0.70	-0.24	0.57	-0.79	0.75	-0.15	0.67	-0.24	0.81	-0.22	0.56	0.08	-0.32	-0.24	-0.10	0.45	-0.00	-0.00	1.00

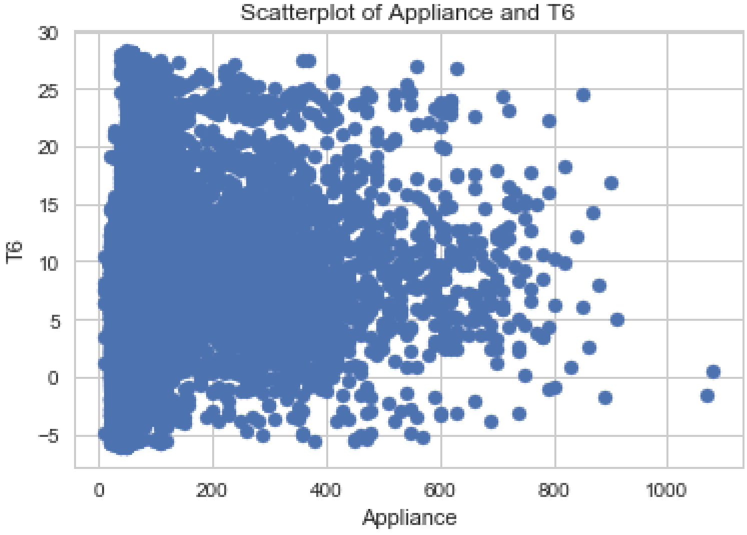
Scatterplot between appliances and t2



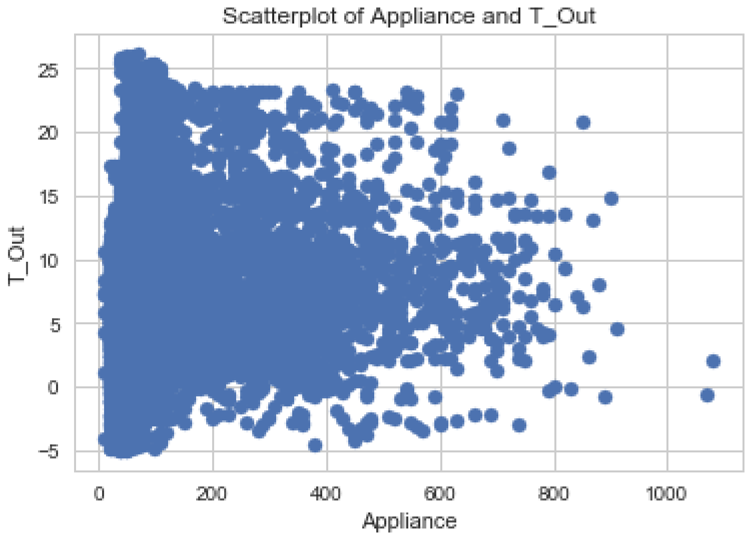
Scatter plot between Appliance and Lights



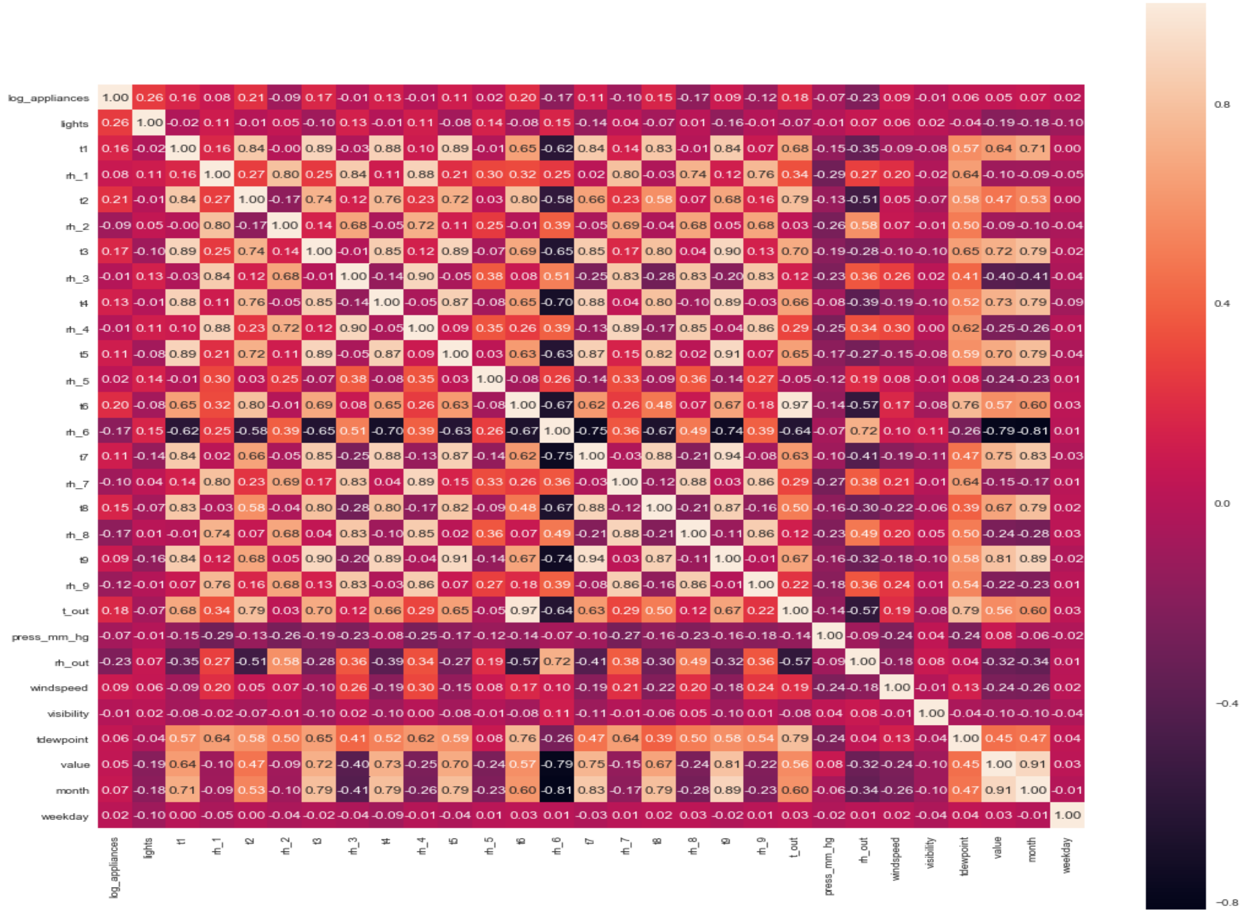
Scatterplot between Appliance and T6



Scatterplot between Appliance and T_out



Correlation plot of using Log value of Appliance -



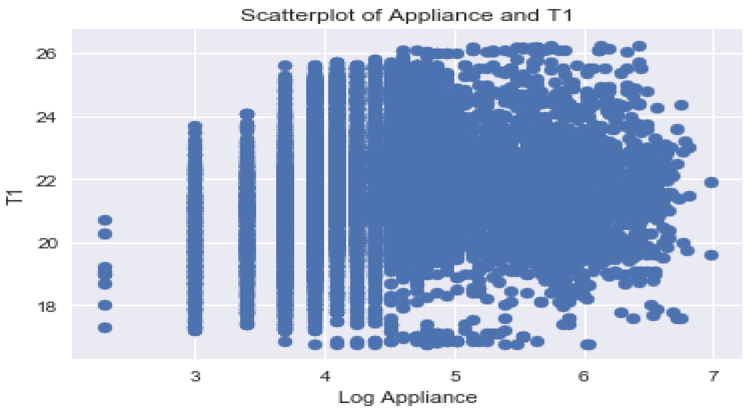
	log_appliances	lights	t1	rh_1	t2	rh_2	t3	rh_3	t4	rh_4	t5	rh_5	t6	rh_6	t7	rh_7	t8	rh_8	t9	rh_9	t_out	press_mm_hg	rh_out	windspeed	visibility	tdewpoint	value	month	weekday
log_appliances	1.00	0.26	0.16	0.08	0.21	-0.09	0.17	-0.01	0.13	-0.01	0.11	0.02	0.20	-0.17	0.11	-0.10	0.15	-0.17	0.09	-0.12	0.18	-0.07	-0.23	0.09	-0.01	0.06	0.05	0.07	0.02
lights	0.26	1.00	-0.02	0.11	-0.01	0.05	-0.10	0.13	-0.01	0.11	-0.08	0.14	-0.08	0.15	-0.14	0.04	-0.07	0.01	-0.16	-0.01	-0.07	-0.01	0.07	0.06	0.02	-0.04	-0.19	-0.18	-0.10
t1	0.16	-0.02	1.00	0.16	0.84	-0.00	0.89	-0.03	0.88	0.10	0.89	-0.01	0.65	-0.62	0.84	0.14	0.83	-0.01	0.84	0.07	0.68	-0.15	-0.35	-0.09	-0.08	0.57	0.64	0.71	0.00
rh_1	0.08	0.11	0.16	1.00	0.27	0.80	0.25	0.84	0.11	0.88	0.21	0.30	0.32	0.25	0.02	0.80	-0.03	0.74	0.12	0.76	0.34	-0.29	0.27	0.20	-0.02	0.64	-0.10	-0.09	-0.05
t2	0.21	-0.01	0.84	0.27	1.00	-0.17	0.74	0.12	0.76	0.23	0.72	0.03	0.80	-0.58	0.66	0.23	0.58	0.07	0.68	0.16	0.79	-0.13	-0.51	0.05	-0.07	0.58	0.47	0.53	0.00
rh_2	-0.09	0.05	-0.00	0.80	-0.17	1.00	0.14	0.68	-0.05	0.72	0.11	0.25	-0.01	0.39	-0.05	0.69	-0.04	0.68	0.05	0.68	0.03	-0.26	0.58	0.07	-0.01	0.50	-0.09	-0.10	-0.04
t3	0.17	-0.10	0.89	0.25	0.74	0.14	1.00	-0.01	0.85	0.12	0.89	-0.07	0.69	-0.65	0.85	0.17	0.80	0.04	0.90	0.13	0.70	-0.19	-0.28	-0.10	-0.10	0.65	0.72	0.79	-0.02
rh_3	-0.01	0.13	-0.03	0.84	0.12	0.68	-0.01	1.00	-0.14	0.90	-0.05	0.38	0.08	0.51	-0.25	0.83	-0.28	0.83	-0.20	0.83	0.12	-0.23	0.36	0.26	0.02	0.41	-0.40	-0.41	-0.04
t4	0.13	-0.01	0.88	0.11	0.76	-0.05	0.85	-0.14	1.00	-0.05	0.87	-0.08	0.65	-0.70	0.88	0.04	0.80	-0.10	0.89	-0.03	0.66	-0.08	-0.39	-0.19	-0.10	0.52	0.73	0.79	-0.09
rh_4	-0.01	0.11	0.10	0.88	0.23	0.72	0.12	0.90	-0.05	1.00	0.09	0.35	0.26	0.39	-0.13	0.89	-0.17	0.85	-0.04	0.86	0.29	-0.25	0.34	0.30	0.00	0.62	-0.25	-0.26	-0.01
t5	0.11	-0.08	0.89	0.21	0.72	0.11	0.89	-0.05	0.87	0.09	1.00	0.03	0.63	-0.63	0.87	0.15	0.82	0.02	0.91	0.07	0.65	-0.17	-0.27	-0.15	-0.08	0.59	0.70	0.79	-0.04
rh_5	0.02	0.14	-0.01	0.30	0.03	0.25	-0.07	0.38	-0.08	0.35	0.03	1.00	-0.08	0.26	-0.14	0.33	-0.09	0.36	-0.14	0.27	-0.05	-0.12	0.19	0.08	-0.01	0.08	-0.24	-0.23	0.01
t6	0.20	-0.08	0.65	0.32	0.80	-0.01	0.69	0.08	0.65	0.26	0.63	-0.08	1.00	-0.67	0.62	0.26	0.48	0.07	0.67	0.18	0.97	-0.14	-0.57	0.17	-0.08	0.76	0.57	0.60	0.03
rh_6	-0.17	0.15	-0.62	0.25	-0.58	0.39	-0.65	0.51	-0.70	0.39	-0.63	0.26	-0.67	1.00	-0.75	0.36	-0.67	0.49	-0.74	0.39	-0.64	-0.07	0.72	0.10	0.11	-0.26	-0.79	-0.81	0.01
t7	0.11	-0.14	0.84	0.02	0.66	-0.05	0.85	-0.25	0.88	-0.13	0.87	-0.14	0.62	-0.75	1.00	-0.03	0.88	-0.21	0.94	-0.08	0.63	-0.10	-0.41	-0.19	-0.11	0.47	0.75	0.83	-0.03
rh_7	-0.10	0.04	0.14	0.80	0.23	0.69	0.17	0.83	0.04	0.89	0.15	0.33	0.26	0.36	-0.03	1.00	-0.12	0.88	0.03	0.86	0.29	-0.27	0.38	0.21	-0.01	0.64	-0.15	-0.17	0.01
t8	0.15	-0.07	0.83	-0.03	0.58	-0.04	0.80	-0.28	0.80	-0.17	0.82	-0.09	0.48	-0.67	0.88	-0.12	1.00	-0.21	0.87	-0.16	0.50	-0.16	-0.30	-0.22	-0.06	0.39	0.67	0.79	0.02
rh_8	-0.17	0.01	-0.01	0.74	0.07	0.68	0.04	0.83	-0.10	0.85	0.02	0.36	0.07	0.49	-0.21	0.88	-0.21	1.00	-0.11	0.86	0.12	-0.23	0.49	0.20	0.05	0.50	-0.24	-0.28	0.03
t9	0.09	-0.16	0.84	0.12	0.68	0.05	0.90	-0.20	0.89	-0.04	0.91	-0.14	0.67	-0.74	0.94	0.03	0.87	-0.11	1.00	-0.01	0.67	-0.16	-0.32	-0.18	-0.10	0.58	0.81	0.89	-0.02
rh_9	-0.12	-0.01	0.07	0.76	0.16	0.68	0.13	0.83	-0.03	0.86	0.07	0.27	0.18	0.39	-0.08	0.86	-0.16	0.86	-0.01	1.00	0.22	-0.18	0.36	0.24	0.01	0.54	-0.22	-0.23	0.01
t_out	0.18	-0.07	0.68	0.34	0.79	0.03	0.70	0.12	0.66	0.29	0.65	-0.05	0.97	-0.64	0.63	0.29	0.50	0.12	0.67	0.22	1.00	-0.14	-0.57	0.19	-0.08	0.79	0.56	0.60	0.03
press_mm_hg	-0.07	-0.01	-0.15	-0.29	-0.13	-0.26	-0.19	-0.23	-0.08	-0.25	-0.17	-0.12	-0.14	-0.07	-0.10	-0.27	-0.16	-0.23	-0.16	-0.18	-0.14	1.00	-0.09	-0.24	0.04	-0.24	0.08	-0.06	-0.02
rh_out	-0.23	0.07	-0.35	0.27	-0.51	0.58	-0.28	0.36	-0.39	0.34	-0.27	0.19	-0.57	0.72	-0.41	0.38	-0.30	0.49	-0.32	0.36	-0.57	-0.09	1.00	-0.18	0.08	0.04	-0.32	-0.34	0.01
windspeed	0.09	0.06	-0.09	0.20	0.05	0.07	-0.10	0.26	-0.19	0.30	-0.15	0.08	0.17	0.10	-0.19	0.21	-0.22	0.20	-0.18	0.24	0.19	-0.24	-0.18	1.00	-0.01	0.13	-0.24	-0.26	0.02
visibility	-0.01	0.02	-0.08	-0.02	-0.07	-0.01	-0.10	0.02	-0.10	0.00	-0.08	-0.01	-0.08	0.11	-0.11	-0.01	-0.06	0.05	-0.10	0.01	-0.08	0.04	0.08	-0.01	1.00	-0.04	-0.10	-0.10	-0.04
tdewpoint	0.06	-0.04	0.57	0.64	0.58	0.50	0.65	0.41	0.52	0.62	0.59	0.08	0.76	-0.26	0.47	0.64	0.39	0.50	0.58	0.54	0.79	-0.24	0.04	0.13	-0.04	1.00	0.45	0.47	0.04
value	0.05	-0.19	0.64	-0.10	0.47	-0.09	0.72	-0.40	0.73	-0.25	0.70	-0.24	0.57	-0.79	0.75	-0.15	0.67	-0.24	0.81	-0.22	0.56	0.08	-0.32	-0.24	-0.10	0.45	1.00	0.91	0.03
month	0.07	-0.18	0.71	-0.09	0.53	-0.10	0.79	-0.41	0.79	-0.26	0.79	-0.23	0.60	-0.81	0.83	-0.17	0.79	-0.28	0.89	-0.23	0.60	-0.06	-0.34	-0.26	-0.10	0.47	0.91	1.00	-0.01
weekday	0.02	-0.10	0.00	-0.05	0.00	-0.04	-0.02	-0.04	-0.09	-0.01	-0.04	0.01	0.03	0.01	-0.03	0.01	0.02	0.03	-0.02	0.01	0.03	-0.02	0.01	0.02	-0.04	0.04	0.03	-0.01	1.00

With Log Appliance

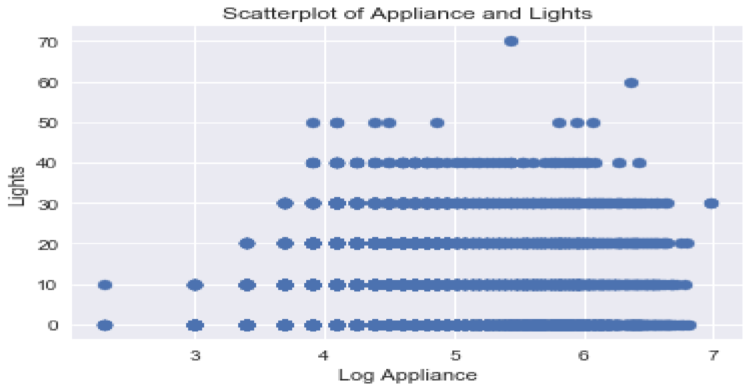
- The most correlated features with energy consumption(log_appliances) are: lights=0.26, t6=0.20, t2=0.22, t3 = 0.17,t_out = 0.18, rh_out = -0.23, rh_8 = -0.17, rh_6 = -0.17, windspeed = 0.09.

- In a linear regression problem only linear independent variables can be be used as features to explain energy consumption otherwise we will have multicollinearity issues.

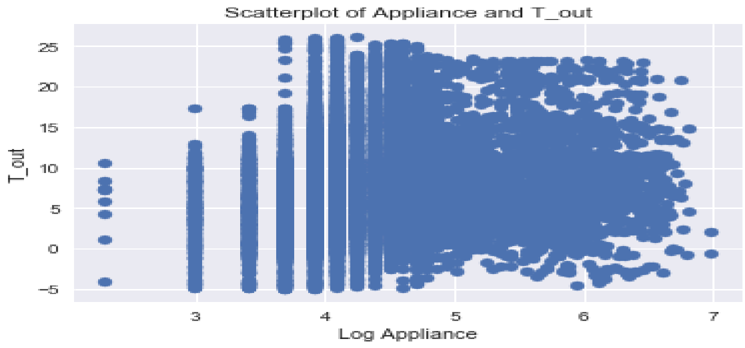
Scatter plot of log_appliance and t1



Scatterplot between log_appliance and lights



Scatter plot between log appliance and t_out


Scatterplot of Appliance and T_out
T_out
Log Appliance