Capstone Project - 2

Supervised machine learning(regression)- Appliance Energy Prediction

Akifuddin Kashif | Zeeshan Ahmed

All about this presentation:

1. Defining problem statement.
2. Overview of data.
3. Performing exploratory data analysis.
4. Model preparation.
5. Building different models.
6. Evaluation of all models.
7. Extracting the best model.

Problem statement

We are tasked with predicting the amount of energy consumed in watt per hour(wh) by tracking the usage of appliances in the household from the data collected throughout 4.5 months at every 10 minute interval to understand the trend and growth of energy consumption of residential buildings in Belgium.

Overview of given data

Date: Date and time of the Appliances usage recorded
Appliances: Values of appliance usage in Watt per hour(Wh)

Lights : Energy use of light fixtures in the house in Wh

T1 : Temperature in kitchen area, in Celsius

RH1, : Humidity in kitchen area, in %

T2: Temperature in living room area, in Celsius

RH2,Humidity in living room area, in %

T3, Temperature in laundry room area

RH3, Humidity in laundry room area, in %

T4, Temperature in office room, in Celsius

RH4,Humidity in office room, in %

T5, Temperature in bathroom, in Celsius

RH5, Humidity in bathroom, in %

T6, Temperature outside the building (north side), in Celsius

RH6, Humidity outside the building (north side), in %

T7, Temperature in ironing room , in Celsius

RH7, Humidity in ironing room, in %

T8, Temperature in teenager room 2, in Celsius

RH8,Humidity in teenager room 2, in %

T9, Temperature in parents room, in Celsius

RH9, Humidity in parents room, in %

To, Temperature outside (from Chievres weather station), in Celsius

Pressure (from Chievres weather station), in mm Hg

RHout, Humidity outside (from Chievres weather station), in %

Wind speed (from Chievres weather station), in m/s

Visibility (from Chievres weather station), in km

Tdewpoint (from Chievres weather station), Â°C

rv1, Random variable 1, nondimensional

rv2, Random variable 2, nondimensional

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19735 entries, 0 to 19734
Data columns (total 29 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   date         19735 non-null  datetime64[ns]
 1   Appliances   19735 non-null  int64
 2   lights       19735 non-null  int64
 3   T1           19735 non-null  float64
 4   RH_1         19735 non-null  float64
 5   T2           19735 non-null  float64
 6   RH_2         19735 non-null  float64
 7   T3           19735 non-null  float64
 8   RH_3         19735 non-null  float64
 9   T4           19735 non-null  float64
 10  RH_4         19735 non-null  float64
 11  T5           19735 non-null  float64
 12  RH_5         19735 non-null  float64
 13  T6           19735 non-null  float64
 14  RH_6         19735 non-null  float64
 15  T7           19735 non-null  float64
 16  RH_7         19735 non-null  float64
 17  T8           19735 non-null  float64
 18  RH_8         19735 non-null  float64
 19  T9           19735 non-null  float64
 20  RH_9         19735 non-null  float64
 21  T_out        19735 non-null  float64
 22  Press_mm_hg  19735 non-null  float64
 23  RH_out       19735 non-null  float64
 24  Windspeed    19735 non-null  float64
 25  Visibility   19735 non-null  float64
 26  Tdewpoint    19735 non-null  float64
 27  rv1          19735 non-null  float64
 28  rv2          19735 non-null  float64
dtypes: datetime64[ns](1), float64(26), int64(2)
memory usage: 4.4 MB
```

Description of data

	Appliances	lights	T1	RH_1	T2	RH_2	T3	RH_3	T4	RH_4	...	T9	RH_9
count	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	...	19735.000000	19735.000000
mean	97.694958	3.801875	21.686571	40.259739	20.341219	40.420420	22.267611	39.242500	20.855335	39.026904	...	19.485828	41.552401
std	102.524891	7.935988	1.606066	3.979299	2.192974	4.069813	2.006111	3.254576	2.042884	4.341321	...	2.014712	4.151497
min	10.000000	0.000000	16.790000	27.023333	16.100000	20.463333	17.200000	28.766667	15.100000	27.660000	...	14.890000	29.166667
25%	50.000000	0.000000	20.760000	37.333333	18.790000	37.900000	20.790000	36.900000	19.530000	35.530000	...	18.000000	38.500000
50%	60.000000	0.000000	21.600000	39.656667	20.000000	40.500000	22.100000	38.530000	20.666667	38.400000	...	19.390000	40.900000
75%	100.000000	0.000000	22.600000	43.066667	21.500000	43.260000	23.290000	41.760000	22.100000	42.156667	...	20.600000	44.338095
max	1080.000000	70.000000	26.260000	63.360000	29.856667	56.026667	29.236000	50.163333	26.200000	51.090000	...	24.500000	53.326667

8 rows × 28 columns

Description of data(contd)

T_out	Press_mm_hg	RH_out	Windspeed	Visibility	Tdewpoint	rv1	rv2
19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000
7.411665	755.522602	79.750418	4.039752	38.330834	3.760707	24.988033	24.988033
5.317409	7.399441	14.901088	2.451221	11.794719	4.194648	14.496634	14.496634
-5.000000	729.300000	24.000000	0.000000	1.000000	-6.600000	0.005322	0.005322
3.666667	750.933333	70.333333	2.000000	29.000000	0.900000	12.497889	12.497889
6.916667	756.100000	83.666667	3.666667	40.000000	3.433333	24.897653	24.897653
10.408333	760.933333	91.666667	5.500000	40.000000	6.566667	37.583769	37.583769
26.100000	772.300000	100.000000	14.000000	66.000000	15.500000	49.996530	49.996530

Sample of data

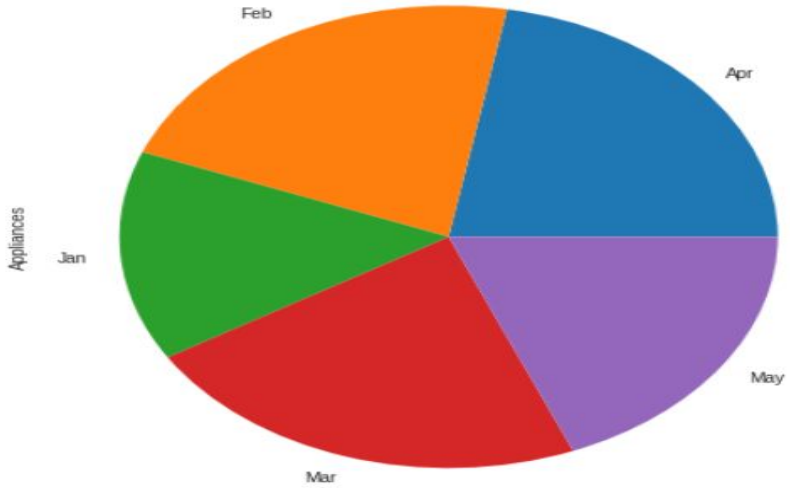
	date	Appliances	lights	T1	RH_1	T2	RH_2	T3	RH_3	T4	...	T9	RH_9	T_out	Press_mm_hg	RH_out	Windspeed	Visibility	Tdewpoint	rv1	rv2
0	2016-01-11 17:00:00	60	30	19.89	47.596667	19.2	44.790000	19.79	44.730000	19.000000	...	17.033333	45.53	6.600000	733.5	92.0	7.000000	63.000000	5.3	13.275433	13.275433
1	2016-01-11 17:10:00	60	30	19.89	46.693333	19.2	44.722500	19.79	44.790000	19.000000	...	17.066667	45.56	6.483333	733.6	92.0	6.666667	59.166667	5.2	18.606195	18.606195
2	2016-01-11 17:20:00	50	30	19.89	46.300000	19.2	44.626667	19.79	44.933333	18.926667	...	17.000000	45.50	6.366667	733.7	92.0	6.333333	55.333333	5.1	28.642668	28.642668
3	2016-01-11 17:30:00	50	40	19.89	46.066667	19.2	44.590000	19.79	45.000000	18.890000	...	17.000000	45.40	6.250000	733.8	92.0	6.000000	51.500000	5.0	45.410389	45.410389
4	2016-01-11 17:40:00	60	40	19.89	46.333333	19.2	44.530000	19.79	45.000000	18.890000	...	17.000000	45.40	6.133333	733.9	92.0	5.666667	47.666667	4.9	10.084097	10.084097

5 rows × 29 columns

Exploratory data analysis

- Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.
- EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

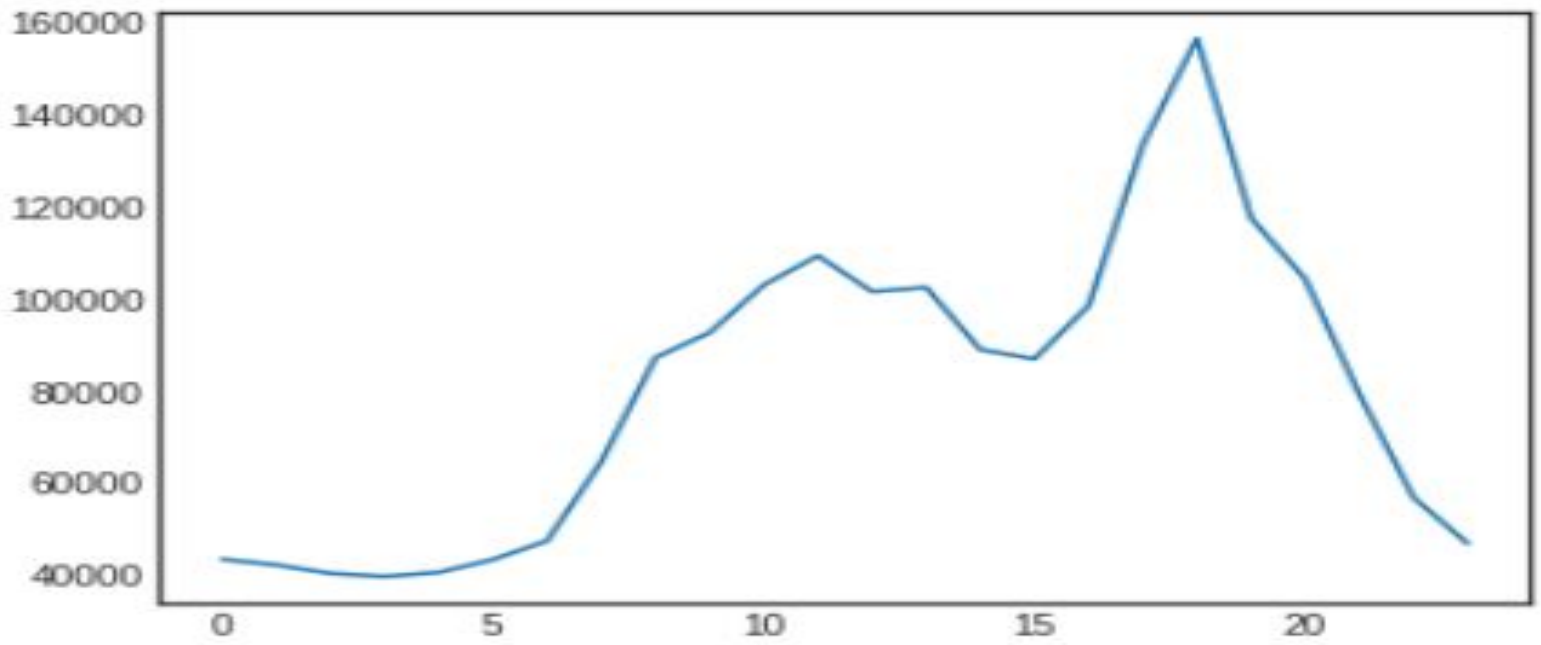
Comparison of Appliances used Seasonally



conclusions from above pie chart:

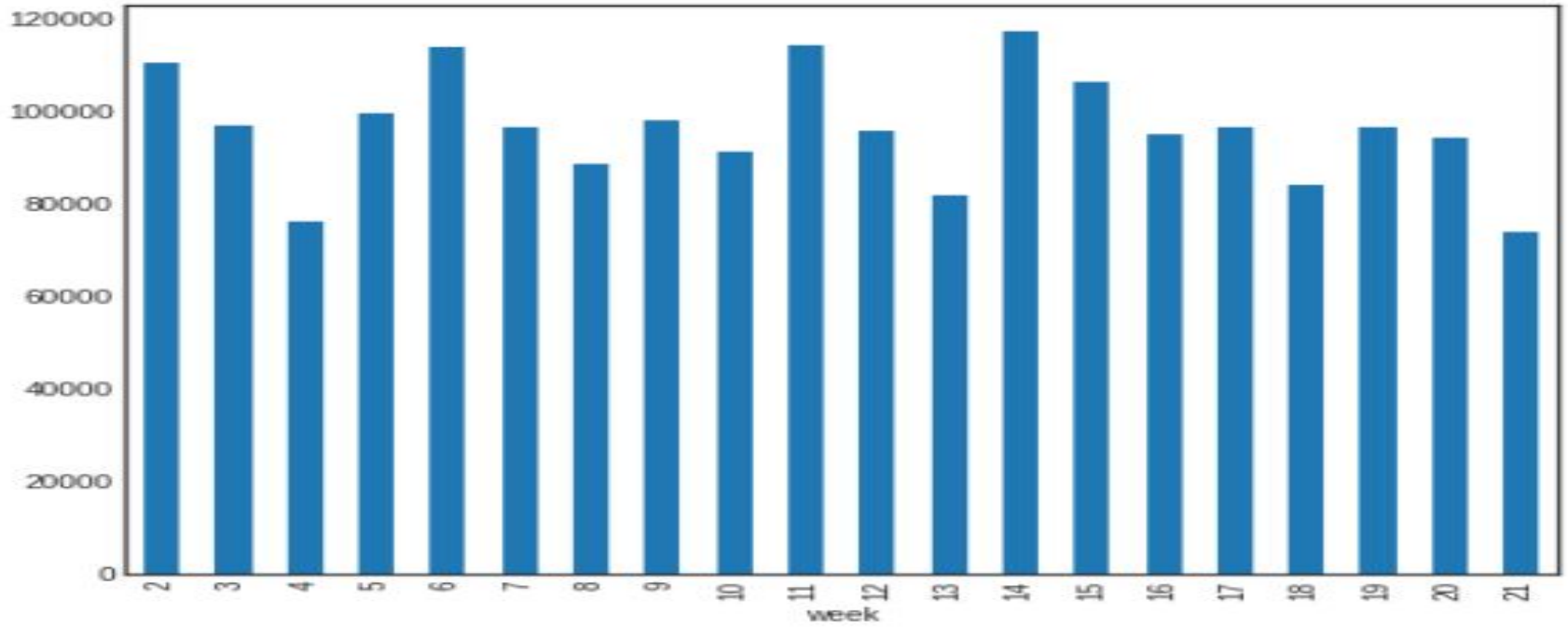
1. In Feb, Mar, April, the usage of appliances is more or less equal.
2. Least appliance usage was recorded in Jan.

Comparison of Hourly Appliance Usage



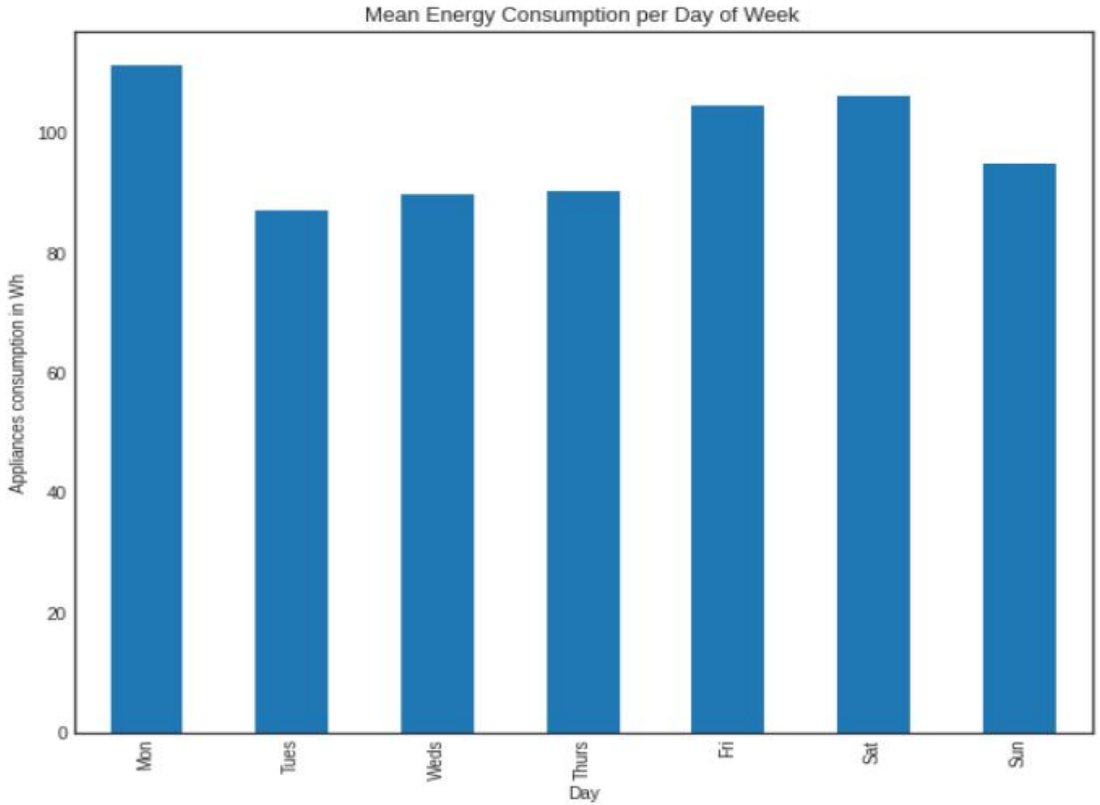
This plot is the sum of all of the energy values across Jan to May for an hour. The usage of appliances is a steady rise starting from 12 am and hits peak aroud 18th hour of the day. Then, after hitting its peak it takes a sudden dip starting from 21st hour.

Comparison of Appliance Usage (all weeks)



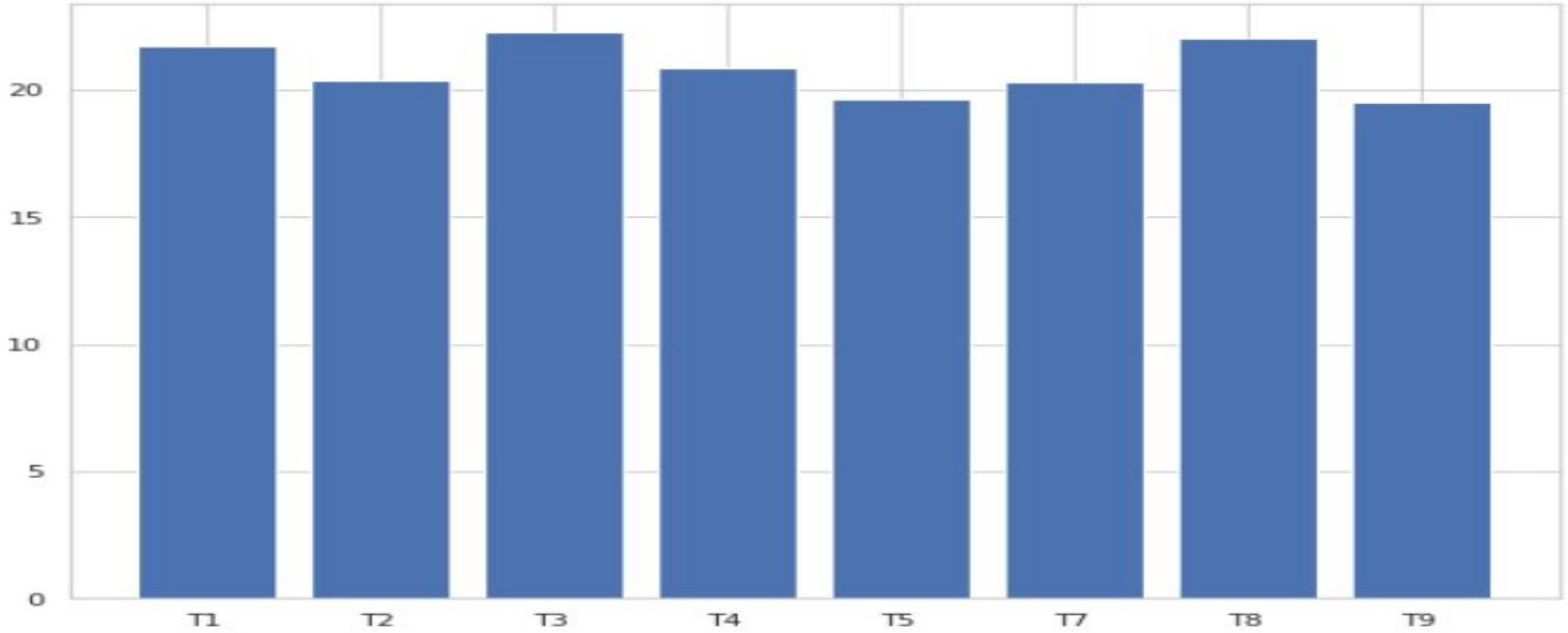
At week 14 i.e, the second week of April, the usage of appliances is the highest and on the other hand, its lowest in the last week of the period i.e, the fourth week of May.

Comparison of Appliance Usage (type of week)



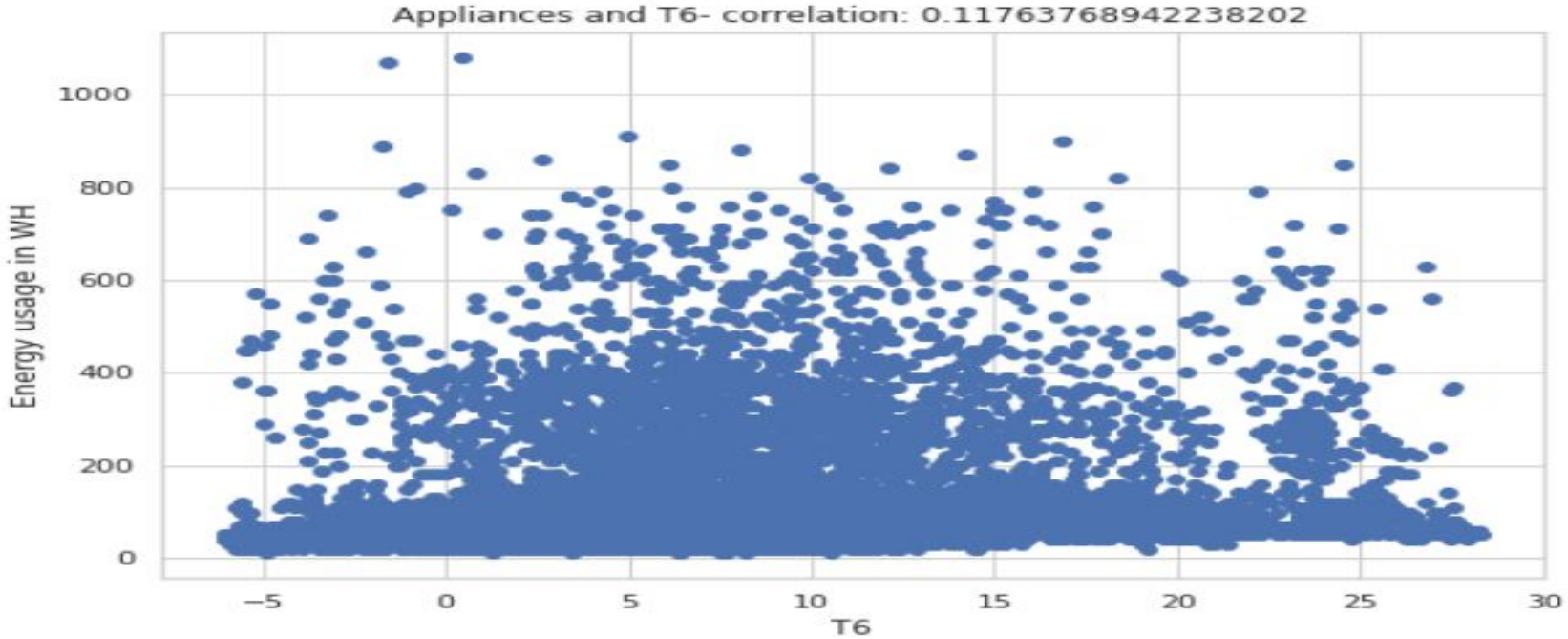
Energy Usage was obeserved to be at its peak on Monday and it was it's listed on Tuesday.

Comparison of variations of temperatures among all the rooms of the building



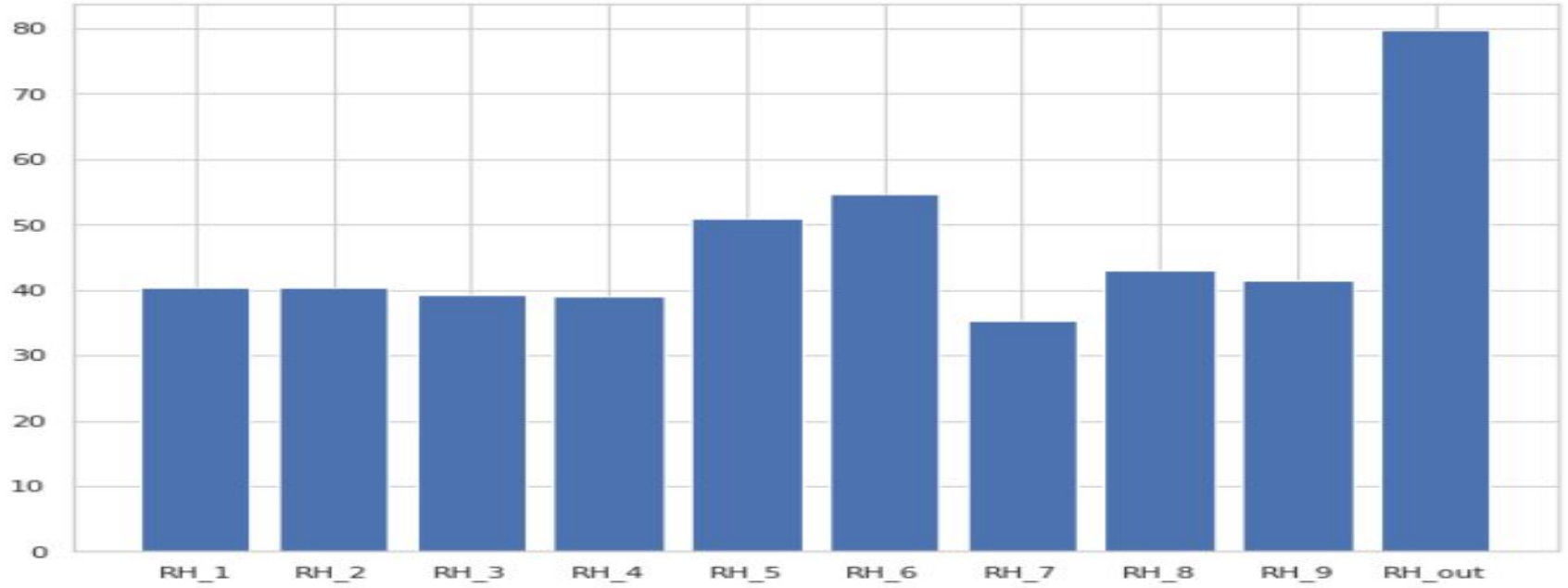
The warmest temperatures are laundry room(T3), teenager room(T8) and kitchen area(T1) respectively. And the coldest rooms are bathroom(T5) and parents room(T9).

Comparison energy variations with respect to temperatures outside the bulding



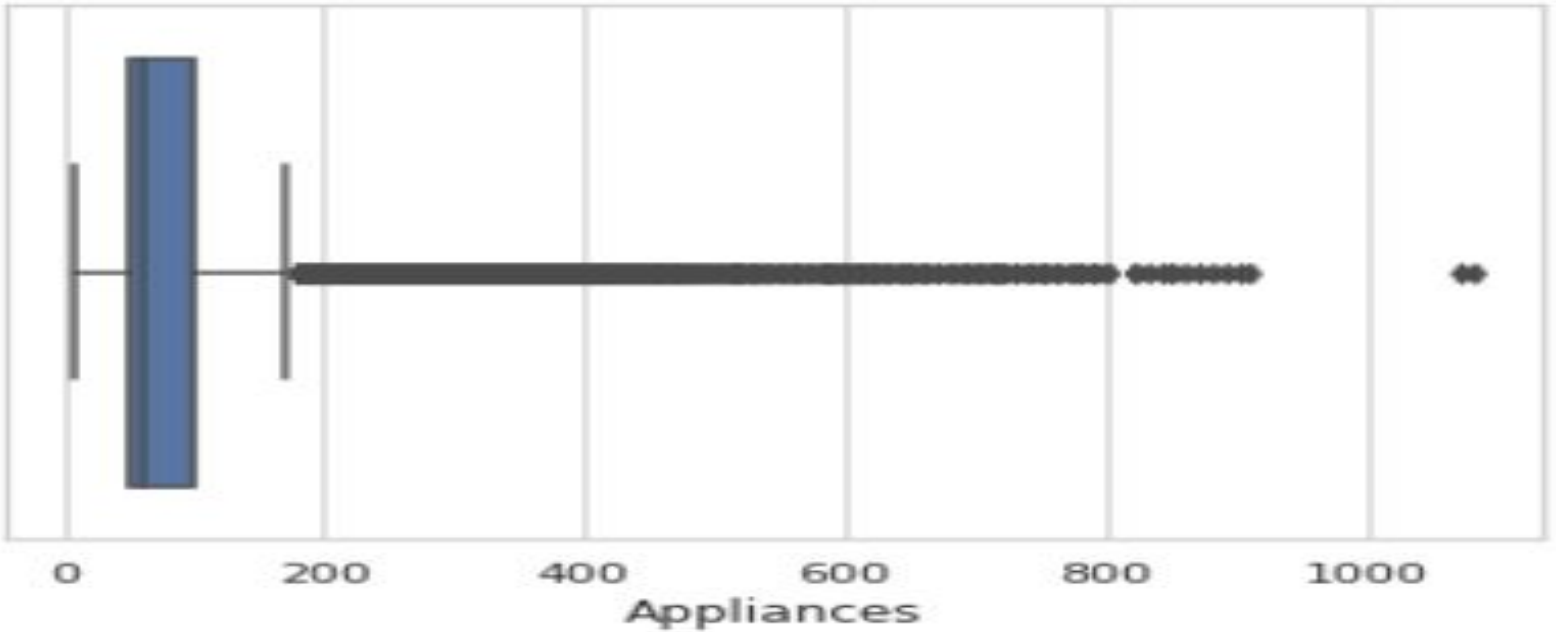
The correlation is minute but it exists between energy use and temperature northside of the building. The energy usage stays consistent from temperatures 5° to 25°. At extreme high and low temperatures, the energy usage is relatively low

Distribution Energy variation with respect to humidity in different rooms of building.



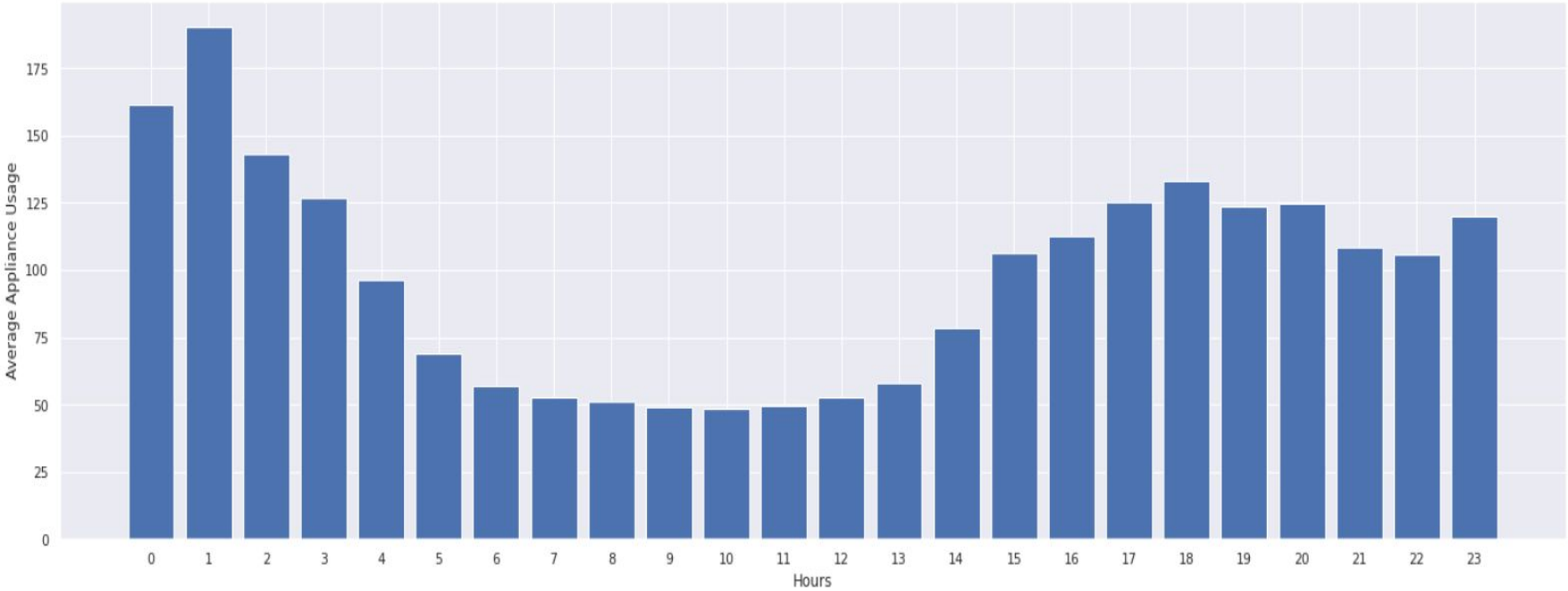
The humidity is at its highest outside the building(RH_out) and lowest in the ironing room(RH_7)

Treating outliers for Appliances column using Box Plot



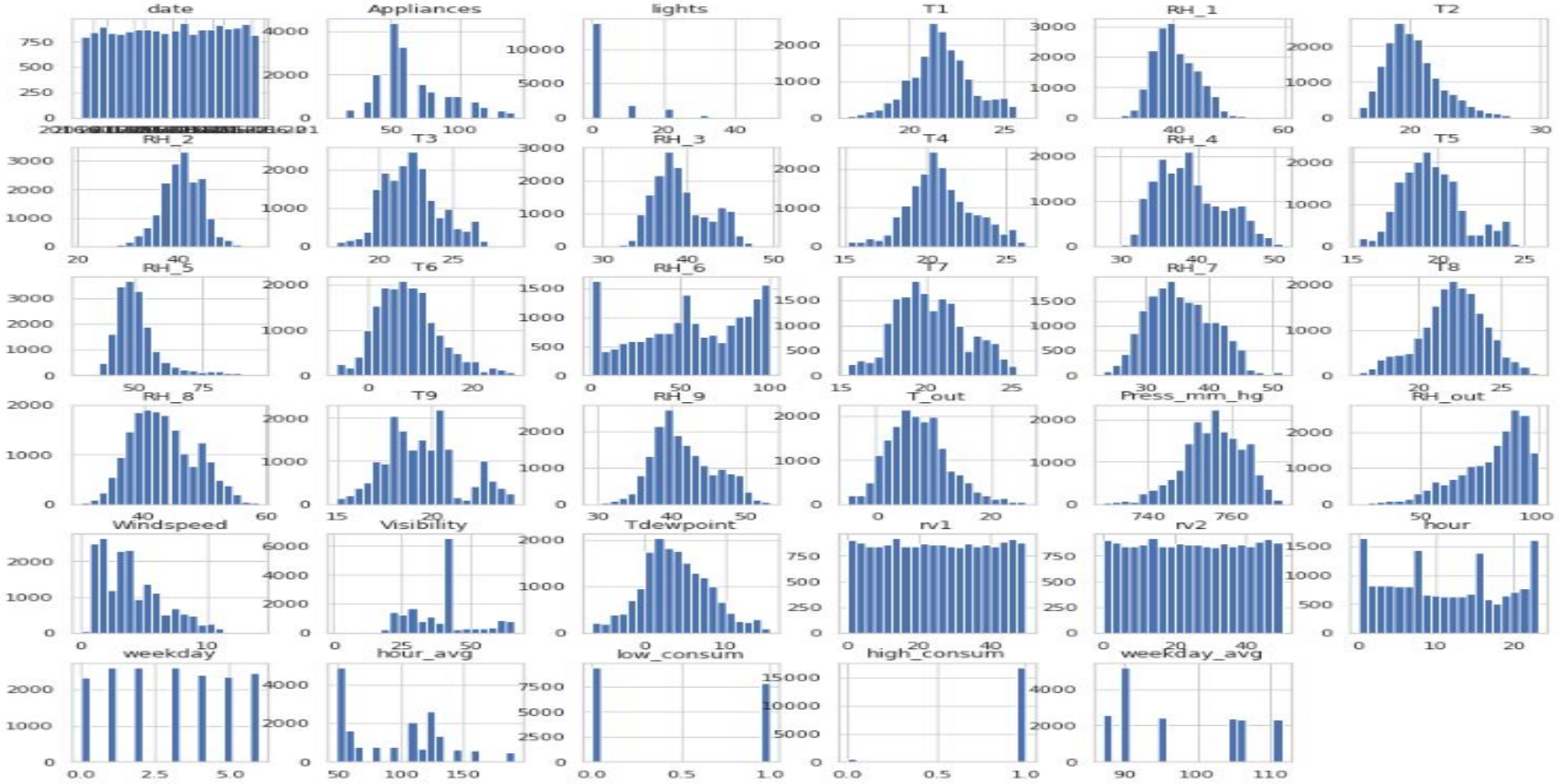
The number of the 0.1% top values of appliances' load is 19 and they have power load higher than 140 Wh as outliers.

Distribution of average use of Appliances



This plot is the average of energy usage values across Jan to May. The energy usage was at its peak during 1st hour and it was as its least during 9th and 10th hour of the day.

Distribution of all dataset column values



Model preparation

1. Calculating multicollinearity through VIF and filtering our data.
2. Plotting the correlation heatmap of all the columns to visualize the collinearity after dropping columns using VIF.
3. All of the values are of desired data type of modelling i.e, Int64 and Float64
4. There aren’t any null values across the dataset.

Correlation Heat Map

This heatmap is plotted after dropping columns which had high VIF scores.

Models used

- Linear regression model
- Lasso regression model
- Ridge regression model
- Decision tree regression model
- Random-forest regression model
- Extra-trees regression model
- XGBoost regression model

Evaluation of models

	model name	R2-score
0	Linear Regression	0.322409
1	Lasso Regression	0.313676
2	Ridge Regression	0.313698
3	Decision Tree Regressor	0.733455
4	Random Forest Regressor	0.791434
5	Extra Trees Regressor	0.798621
6	XGBoost Regressor	0.767874

From above it is clear that extra-trees regression model has done very well with our dataset with an R2- score of ~80% accuracy

Challenges faced

1. Pre-processing the data was one of the challenges we faced which includes removing highly correlated variables from the data so as to not hinder the performance of our regression model.
2. Exploring all the columns and calculating VIF for multicollinearity was challenging because it might decrease the models performance.
3. Selecting the appropriate models to maximize the accuracy of our predictions was one of the challenges faced.

Conclusion

We are finally at the conclusion of our project!

Coming from the beginning we did EDA on the dataset and also cleaned the data according to our needs.After that we were able to draw relevant conclusions from the given data and then we trained our model on linear regression and other models .

Out of all models used , with extra-trees regression model we were able to get the r2-score of 0.80.The model which performed poorly was Lasso regularization with r2-score of 0.31.

Given the size of data and the amount of irrelevance in the data , the above score is good.