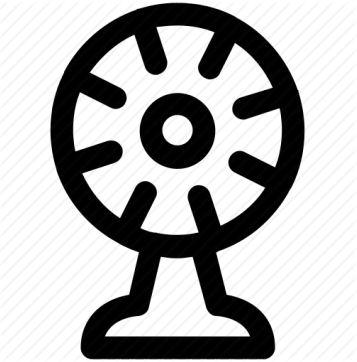
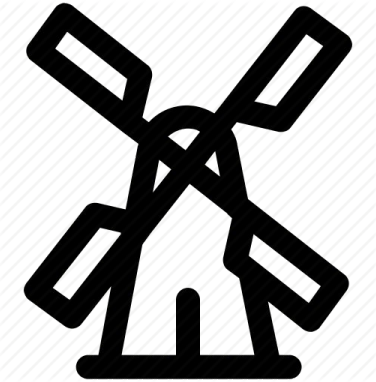
# Capstone Project
# Appliances Energy Prediction





# Problem Statement

**Goal :**

- The goal of our project is to predict the energy consumption of appliances in households based on the sensor data we have from a particular apartment and corresponding weather reports.

**Why is this important ?**

- We need to find out the energy consumption of households in the city so as to ensure that the amount of electricity fed into the electricity grid is always be equal to the amount of electricity consumed, otherwise there's a possibility of black out.

**How is that?**

- If too much electricity is fed into the grid in relation to the quantity consumed, the electrical frequency increases. Since power plants are designed to operate within a certain frequency range, there is a risk that they will disconnect from the grid after a period of time.
- If we feed in too little to meet demand, the frequency drops. From 49 Hz, the automatic load shedding plan is activated in order to avoid power cuts. This is because, if the frequency falls too much, the power plants switch off one after another, until there is a complete collapse of the grid, i.e. a power blackout.

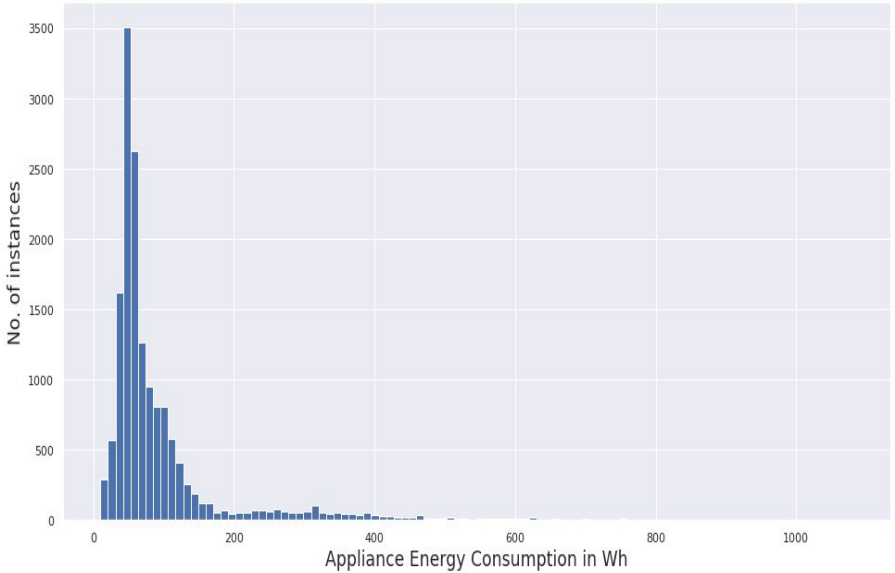
# Data Summary

- The dataset we have is a series of sensor data collected from a building in Belgium at an interval of 10 mins for a period of about 4.5 months.
- The sensor data consists of temperature and humidity levels in different rooms in the building.
- Among other features are weather reports on Pressure, Wind speed, Visibility and T-dew point, which are recorded at weather station Chievres Airport, Belgium .
- The target variable is the total energy consumption of the building in Wh.
- The dataset has no null values.

# Exploratory Data Analysis

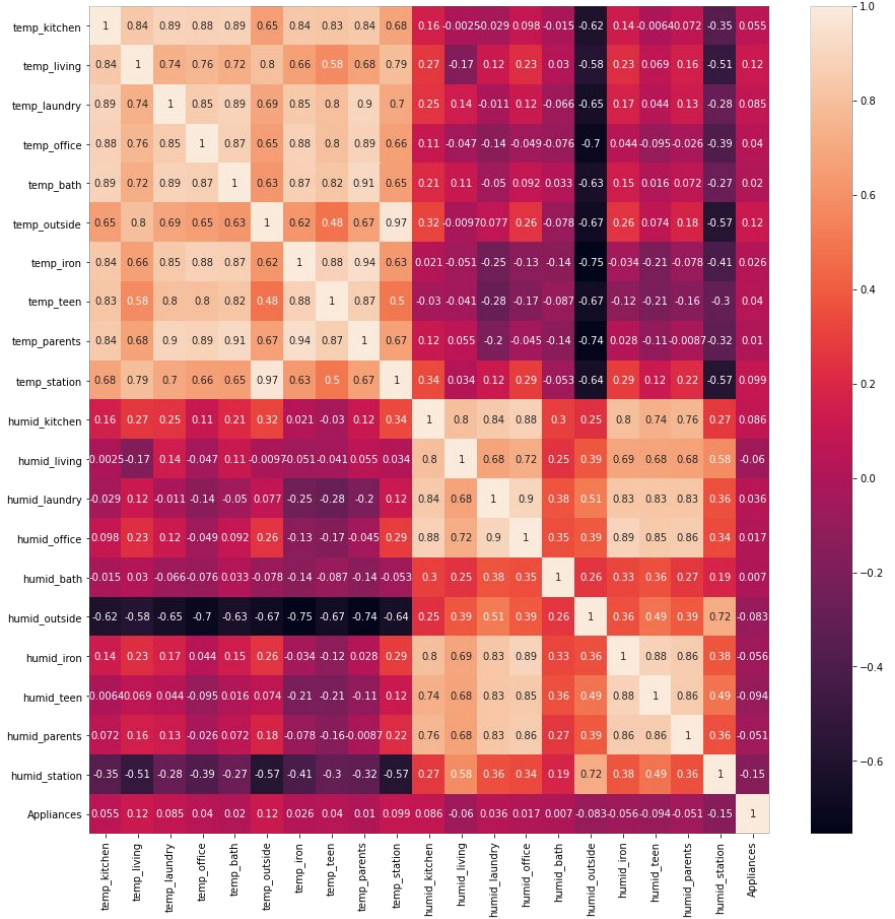
# Target Variable

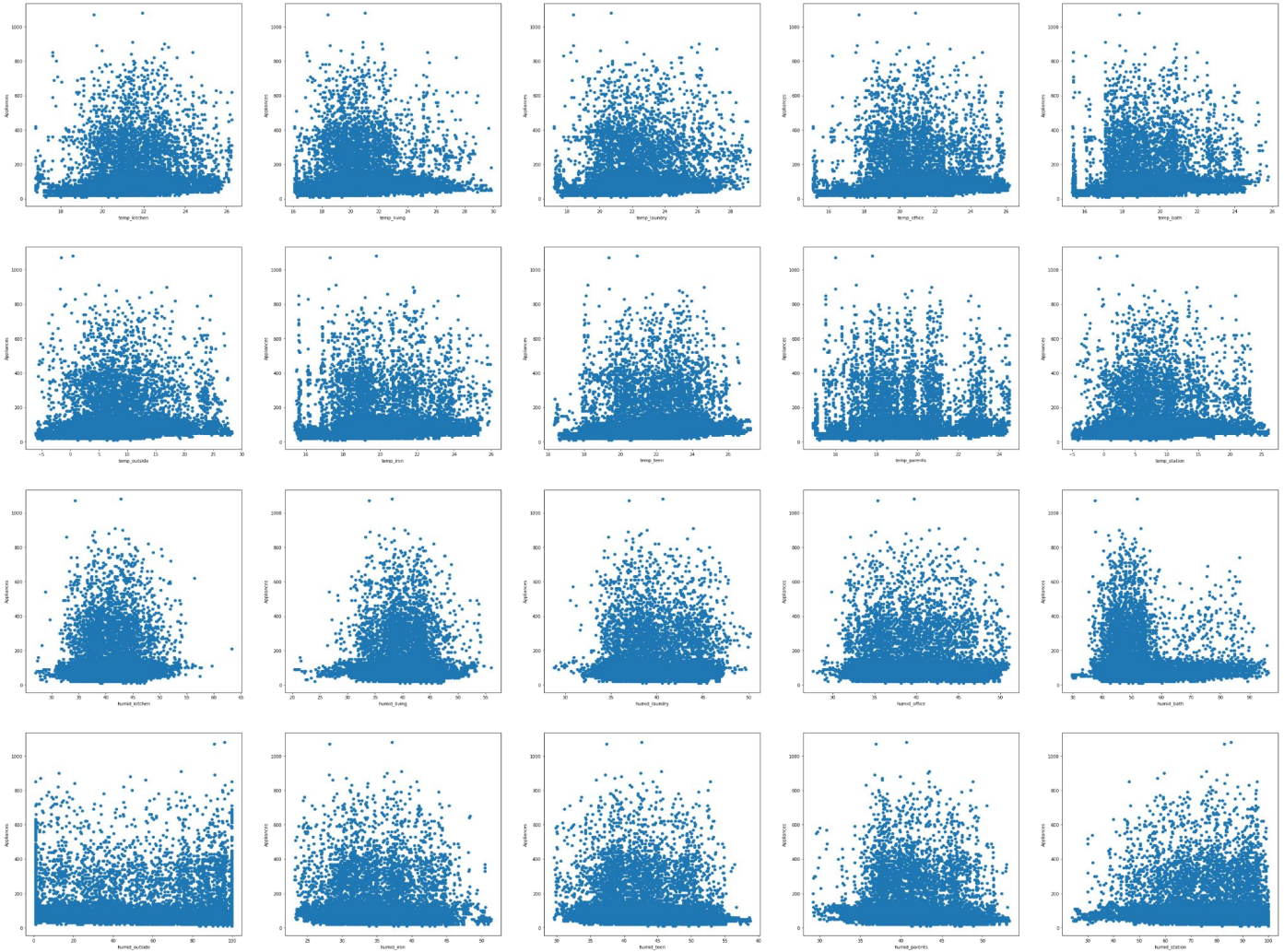
- Energy consumption of appliances ranges from 10 Wh to 1080 Wh
- About 75 % of energy values lie below 100 Wh, and about 93 % of them lie below 300 Wh
- Our target variable seems to be highly skewed, and our task is to predict the usual as well as the large surges in energy in the building.

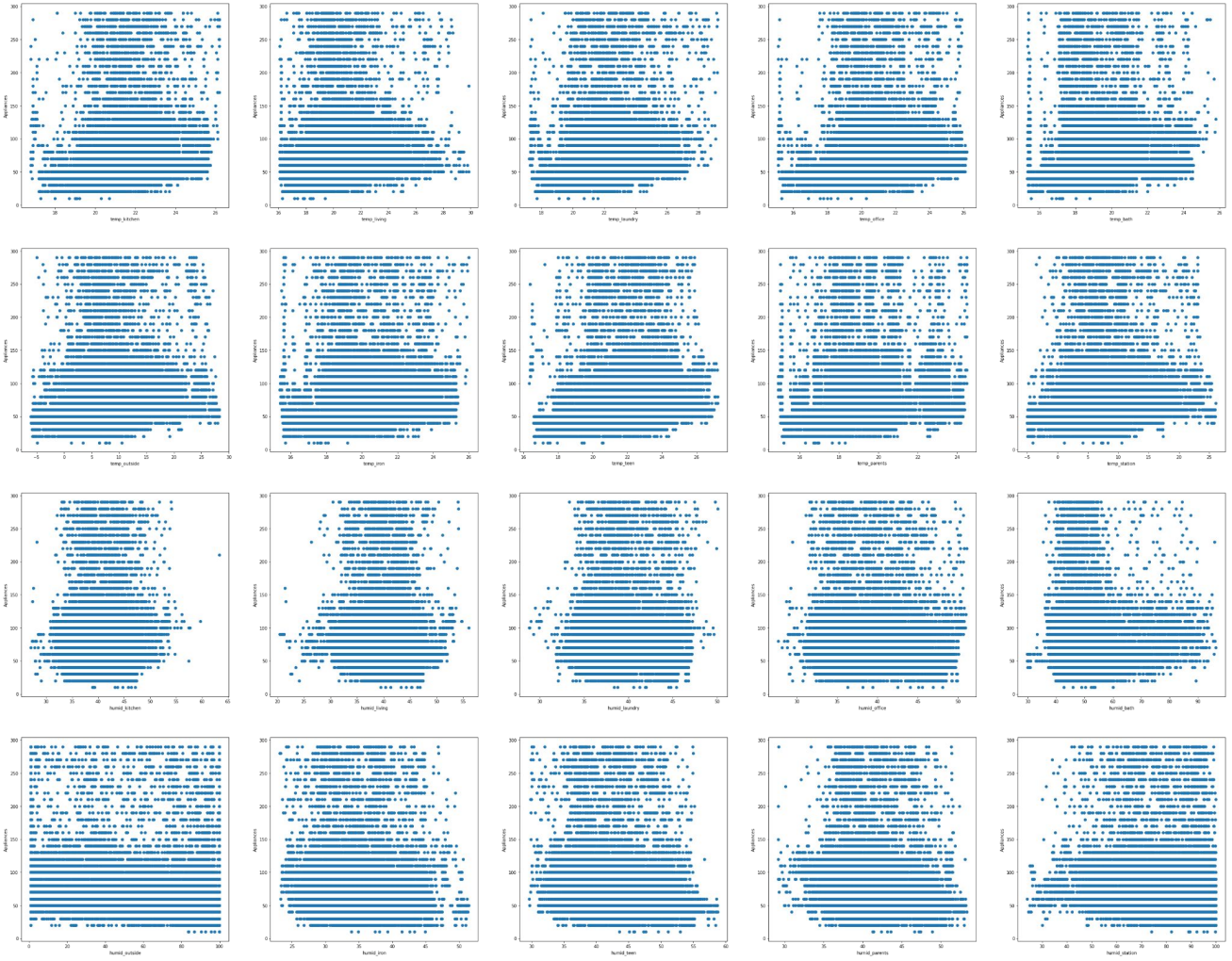


# Correlation matrix

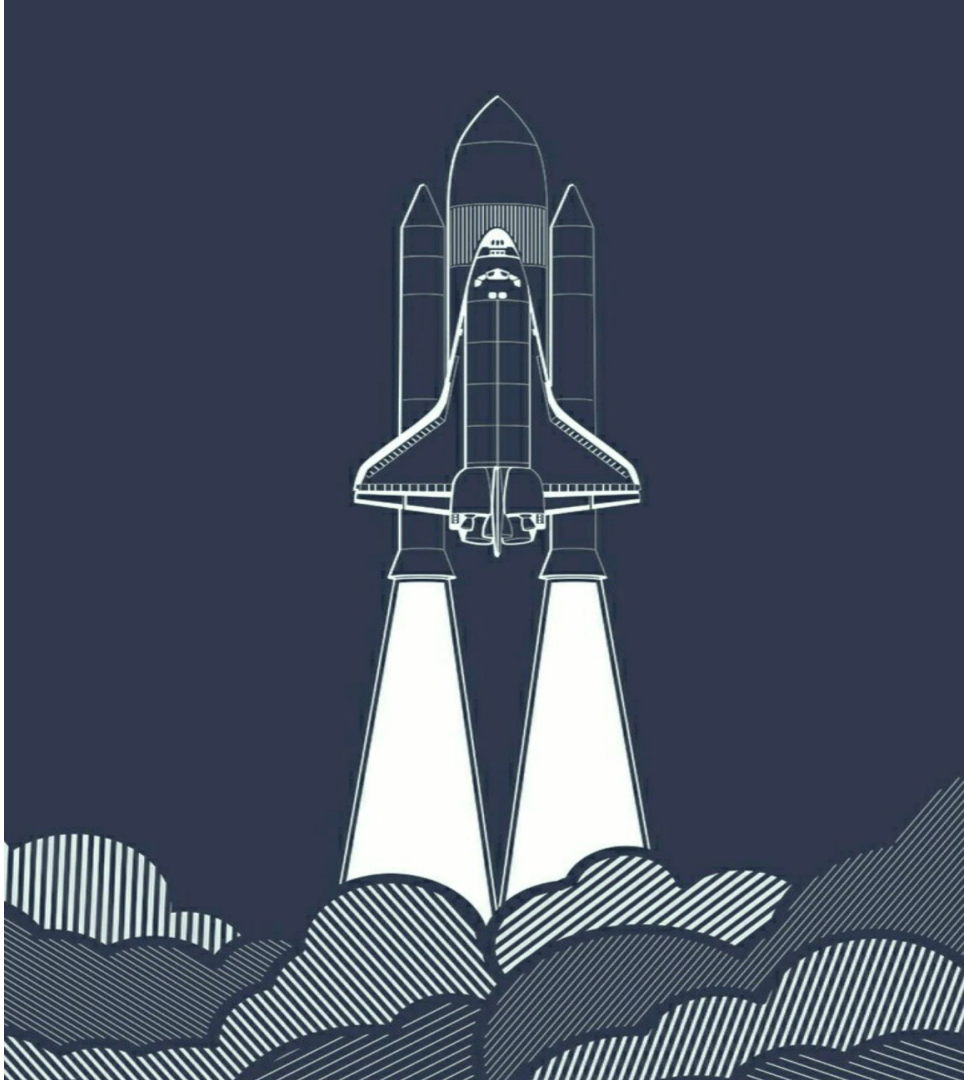
- The Temperature and humidity levels in each of the rooms are highly correlated among themselves.
- There seems to be no relationship between humidity and temperature levels in the building.
- However temperature and humidity levels outside the building are hugely negatively correlated.
- There is little to no correlation between these features and the target variable i.e. Appliance energy consumption




Appliances
temp_kitchen
temp_living
temp_laundry
temp_office
temp_bath
temp_outside
temp_iron
temp_teen
temp_parents
temp_station
humid_kitchen
humid_living
humid_laundry
humid_office
humid_bath
humid_outside
humid_iron
humid_teen
humid_parents
humid_station


Appliances
temp_kitchen
temp_living
temp_laundry
temp_office
temp_bath
temp_outside
temp_iron
temp_teen
temp_parents
temp_station
humid_kitchen
humid_living
humid_laundry
humid_office
humid_bath
humid_outside
humid_iron
humid_teen
humid_parents
humid_station

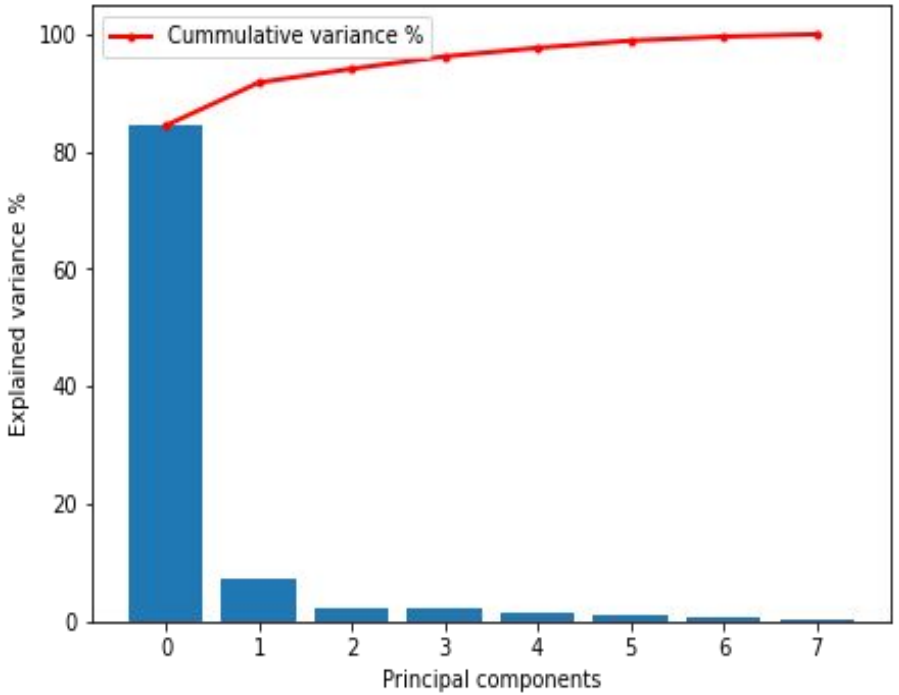
# Feature Engineering



# Principal component Analysis

- Given, temperature levels in different rooms had a very low correlation with target variable, and high correlation among themselves, we reduce the feature set into lower dimensions that could explain maximum variance.

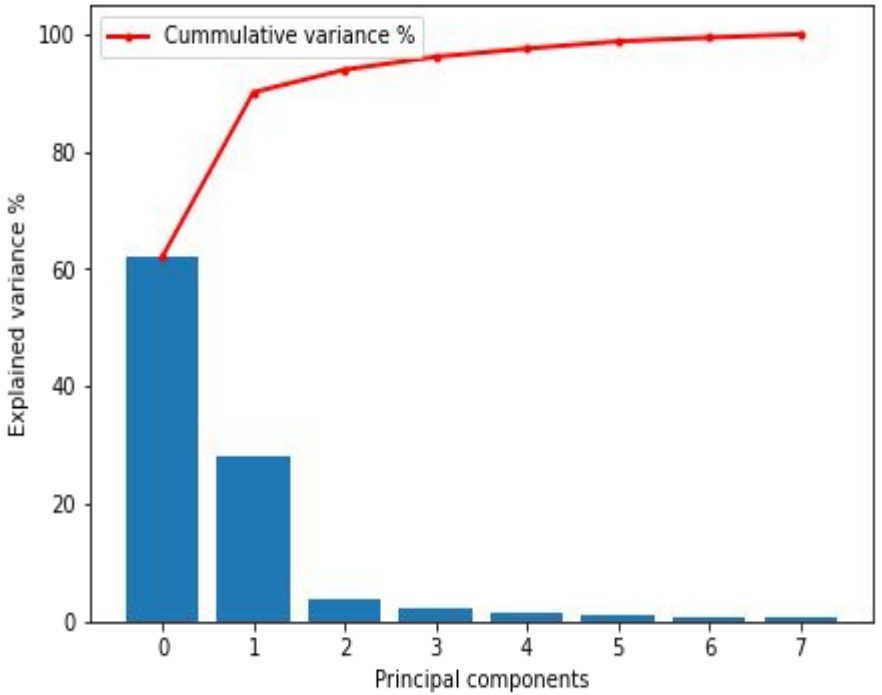
- PCA 1 and 2 explain more than 91 % variance in the temperature levels in different rooms in the building



# Principal component Analysis

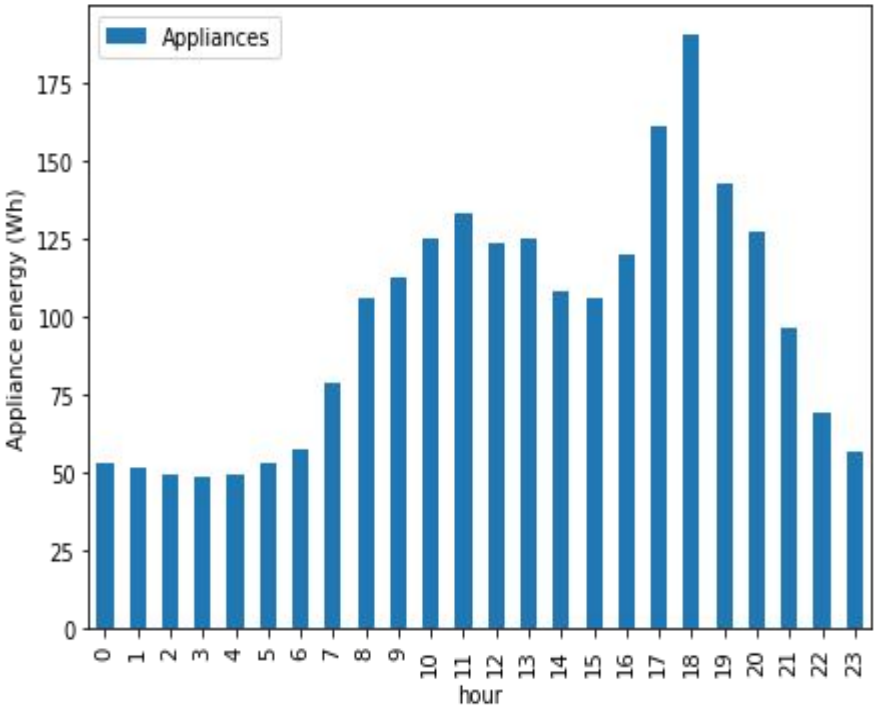
- Given, humidity levels in different rooms had a very low correlation with target variable, and high correlation among themselves, we reduce the feature set into lower dimensions that could explain maximum variance.

- PCA 1 and 2 explain more than 91 % variance in the humidity levels in different rooms in the building

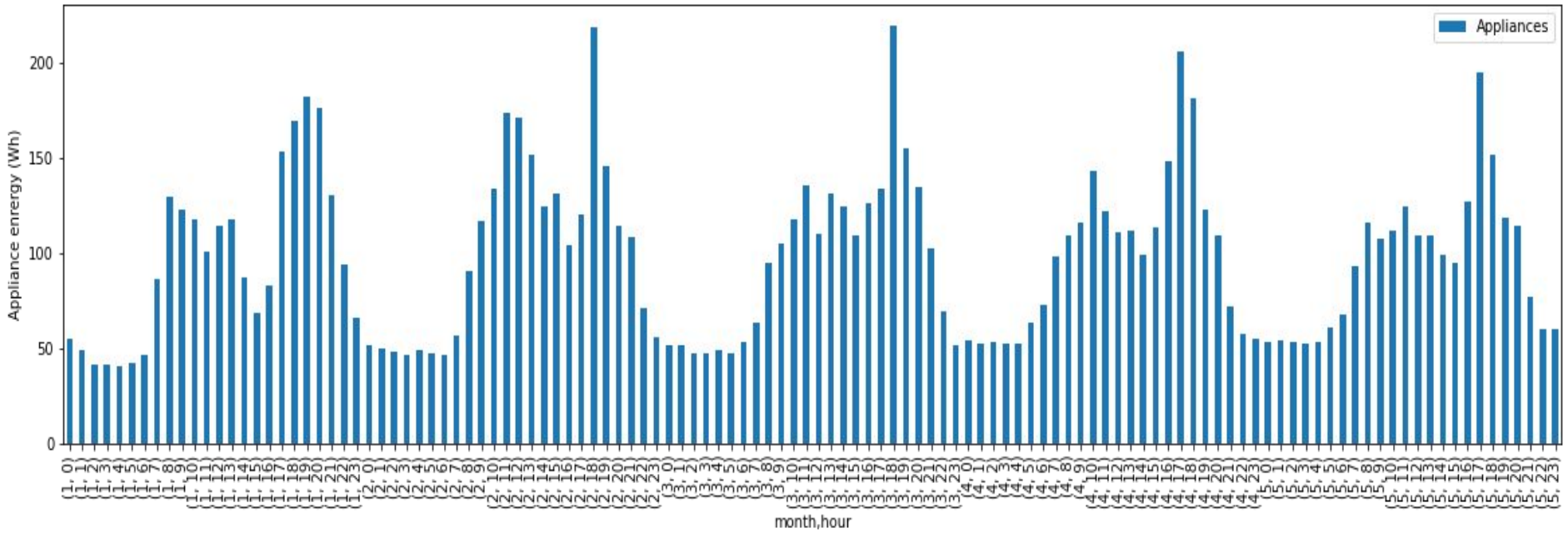


# Date time (Hour)

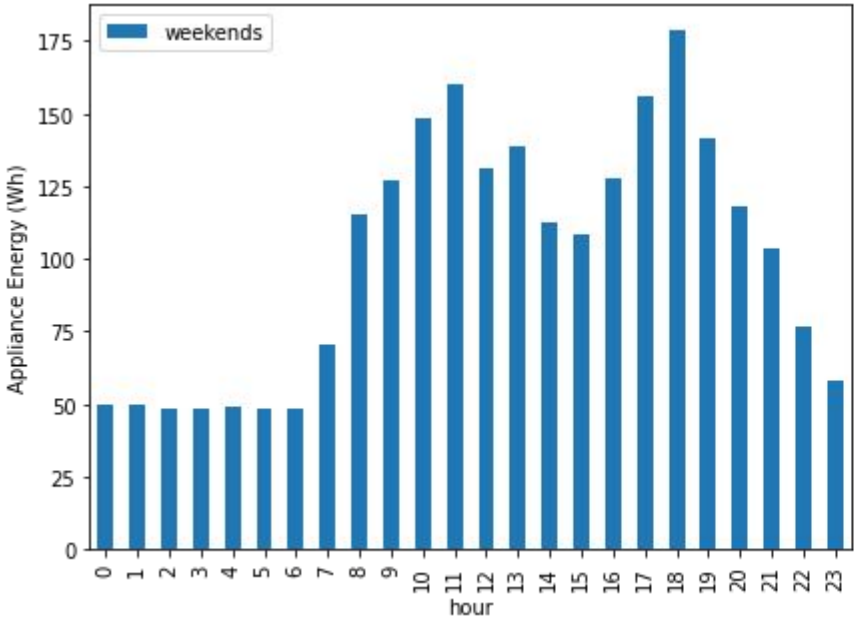
- We have column called date which includes the timestamp corresponding to each of the sensor data sample.
- There seems to be a pattern to energy consumption of appliances at different time of the day.
- Morning 11 AM and evening 6 PM seem to be peak hours.

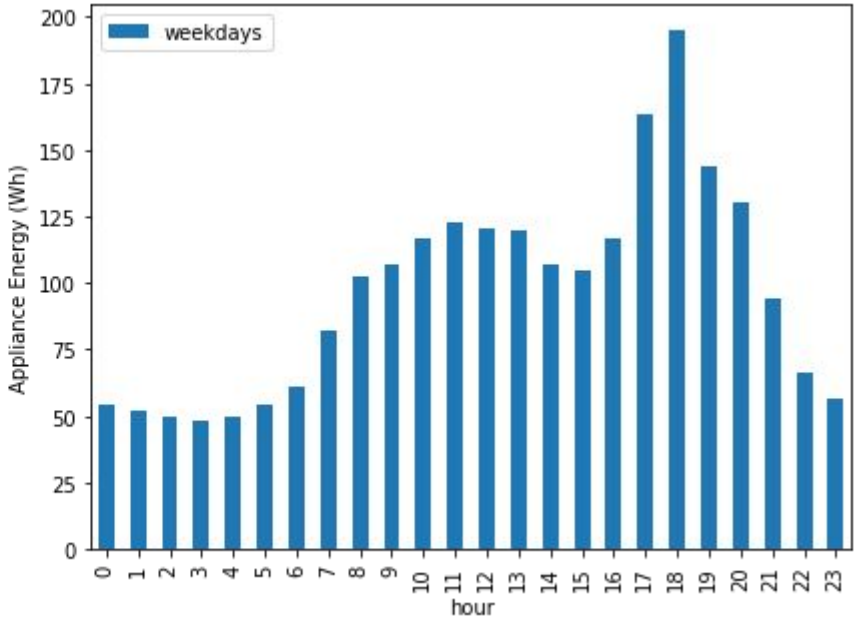


# Date time (Month)



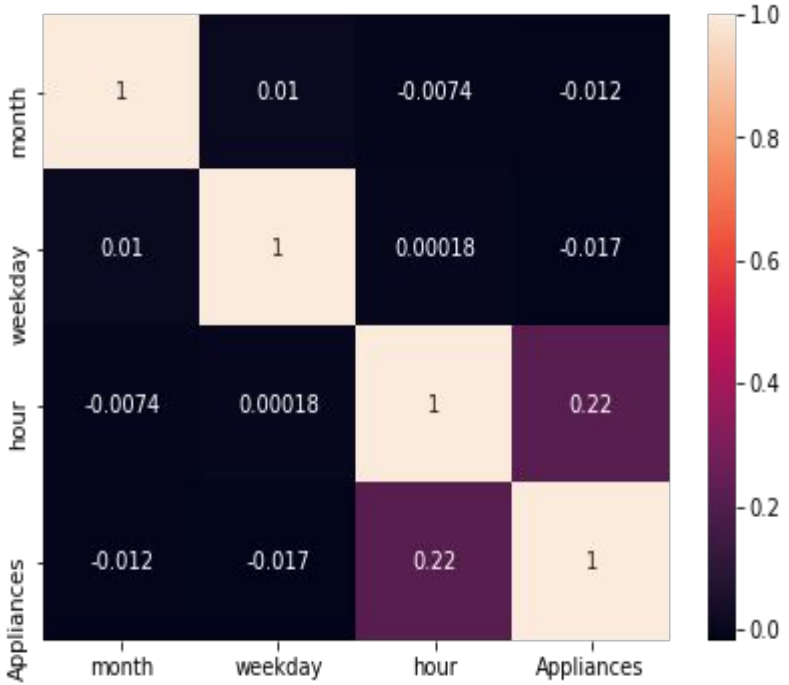
# Date time (Weekday)



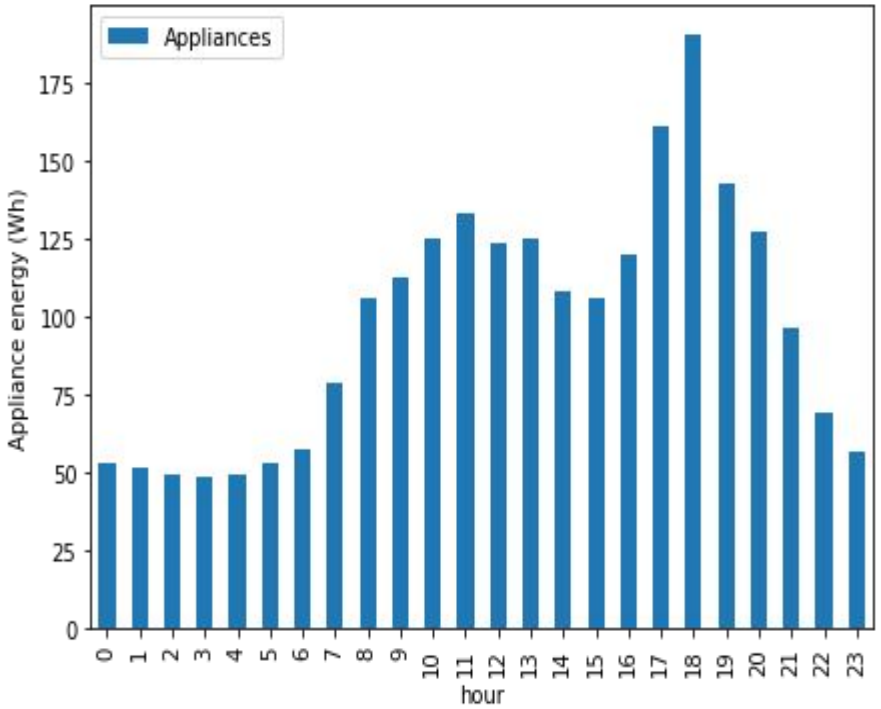
# Correlation

- There is no significant correlation between month and the appliance energy consumption, as we saw the pattern of consumption was almost similar over all months.
- Same is the case for weekdays and weekends.
- However, there is a significant correlation between hours and the target variable.

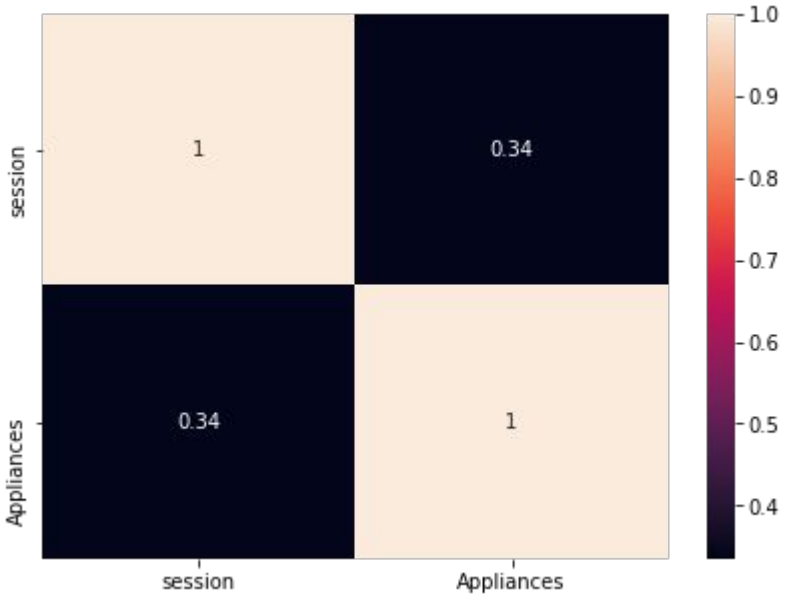


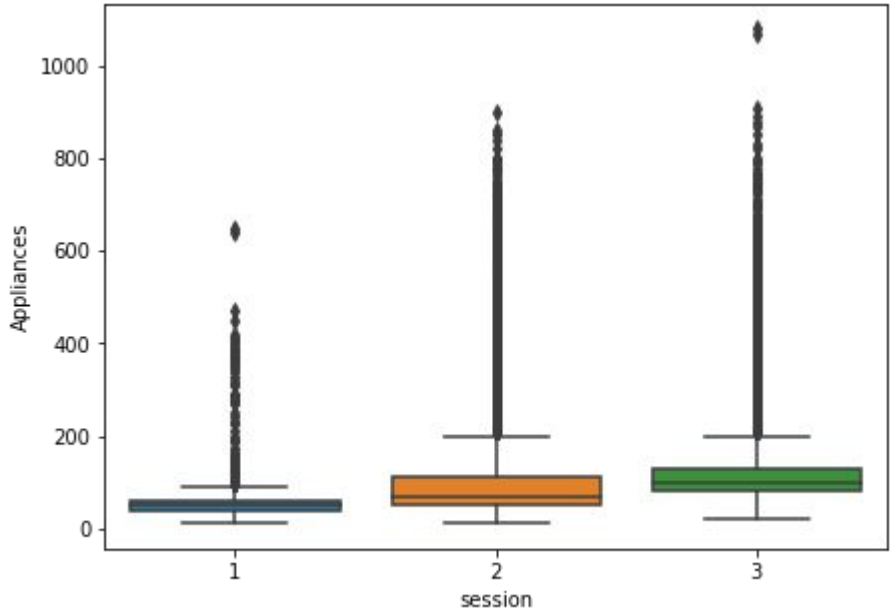
# Session

- We divided the dataset into 3 sessions :
- Class 1 : 10 PM - 6 AM
- Class 2 : 6 AM - 3 PM
- Class 3 : 3 PM - 10 PM



# Session


1
0.34
0.34
1
session
Appliances
session
Appliances
1.0
0.9
0.8
0.7
0.6
0.5
0.4


Appliances
1000
800
600
400
200
0
1
2
3
session

# Final Features

Feature set - I

- Session
- Temperature - pca 1
- Temperature - pca 2
- Humidity - pca 1
- Humidity - pca 2
- Temperature outside
- Humidity outside
- Pressure
- Wind speed

Feature set - II

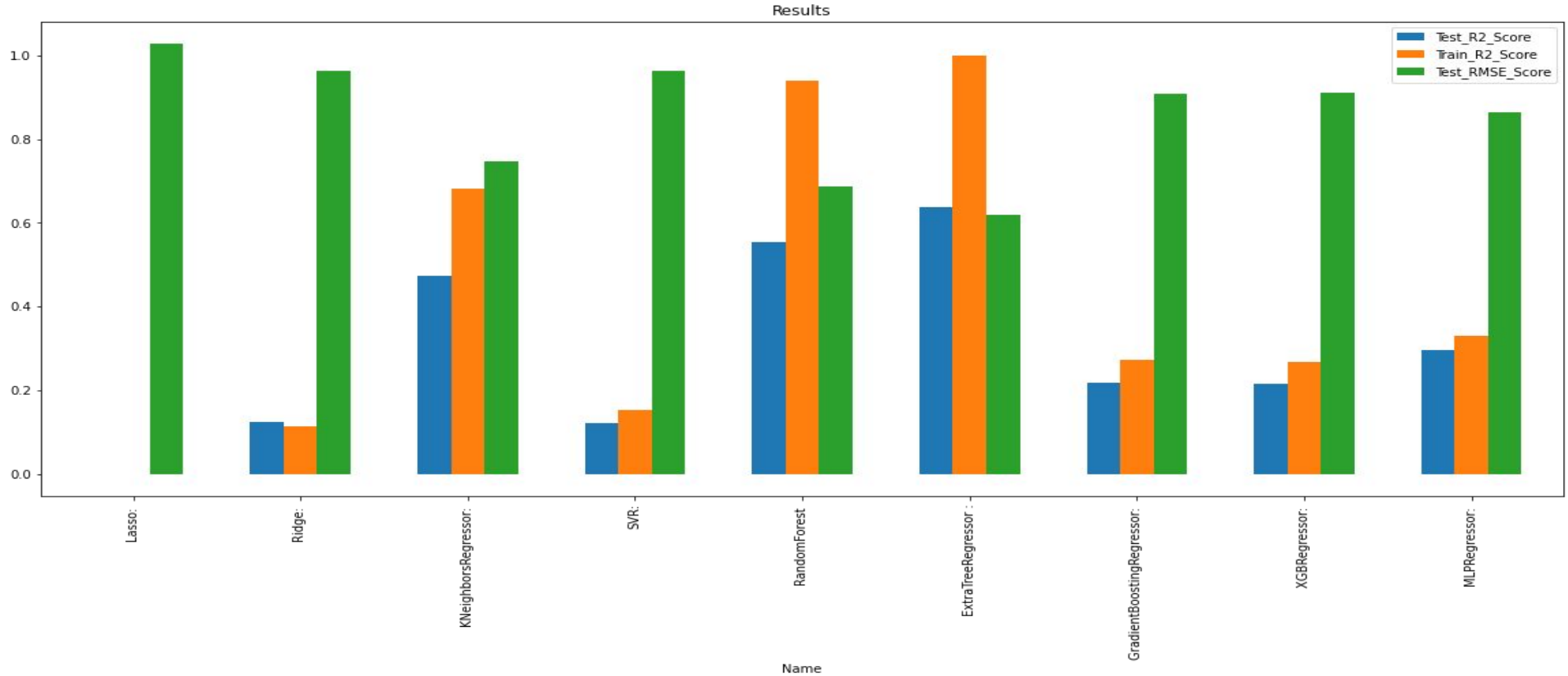
- Session
- Temperature of all rooms inside building
- Humidity of all rooms inside building
- Temperature outside
- Humidity outside
- Pressure
- Wind speed

# Model Training and Evaluation

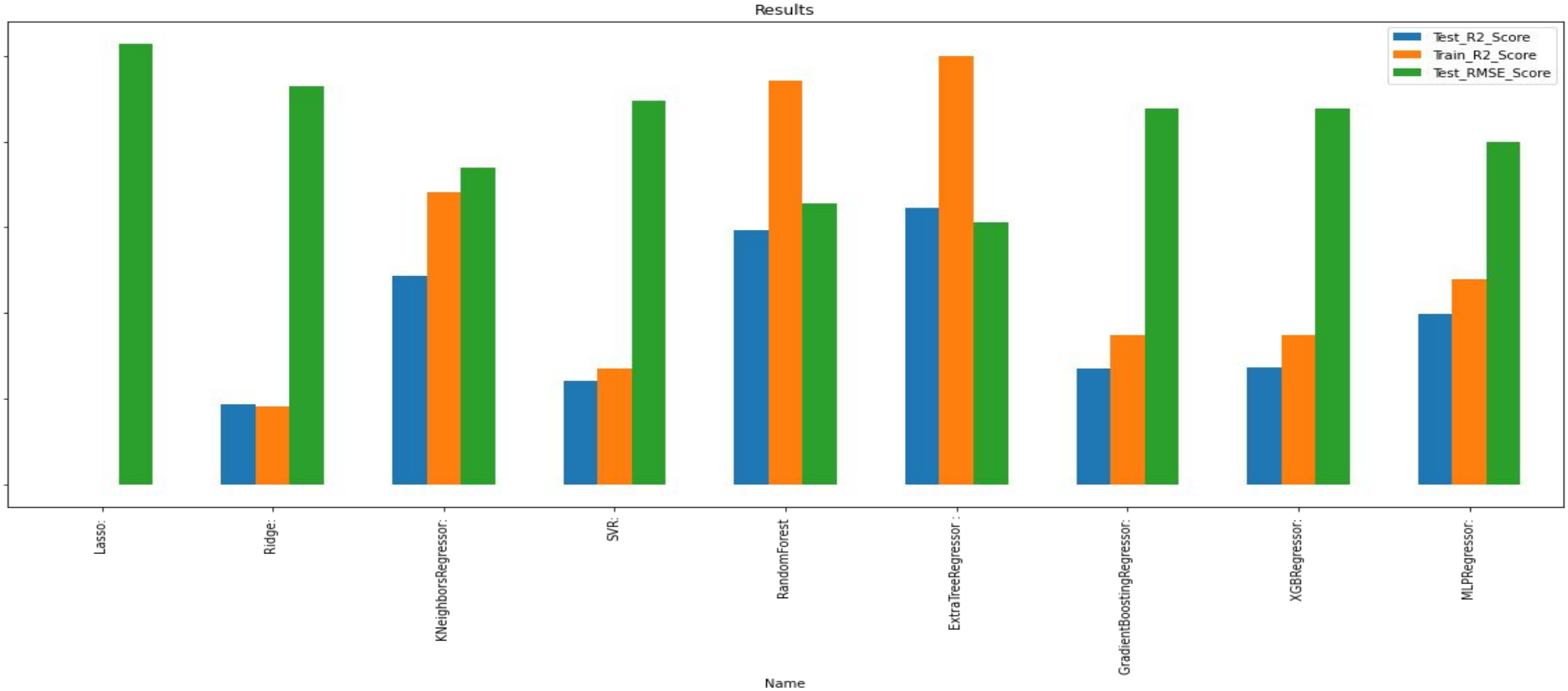
# Models trained

- Lasso and Ridge regression
- Random Forest Regressor
- Extra trees Regressor
- Gradient boosting regressor
- XG-Boost Regressor
- K Neighbours Regressor
- Support Vector Regressor
- MLP regressor (Neural networks)

# Results (with PCA features)



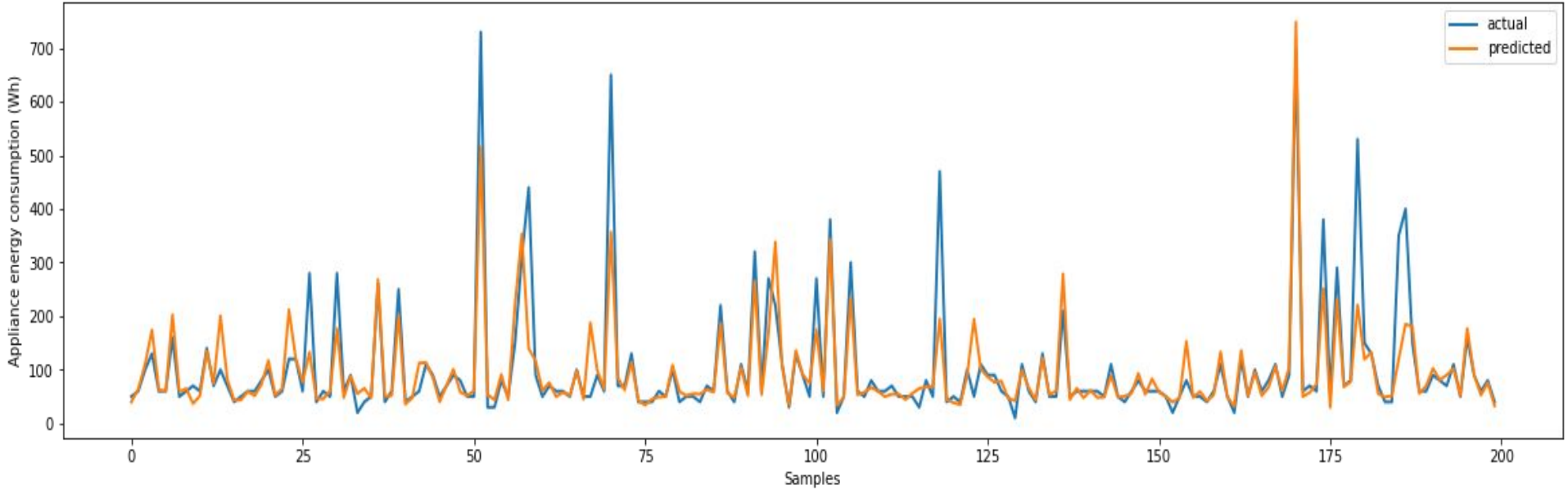
# Results (without PCA features)



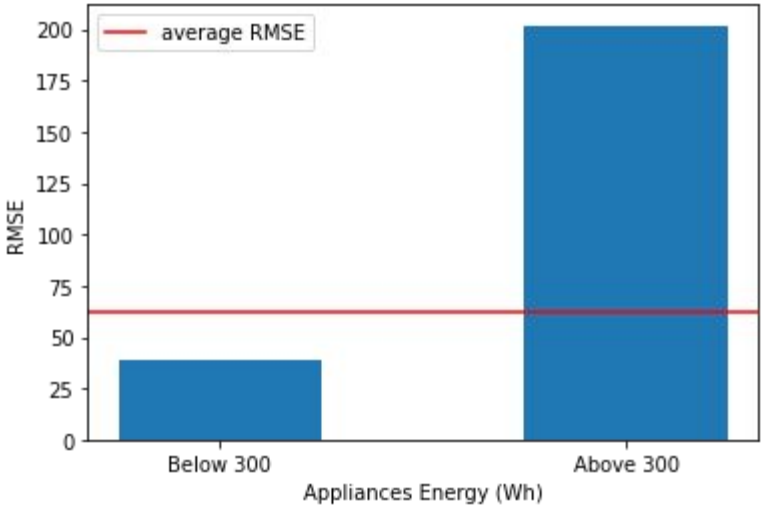
# Best Performing Model

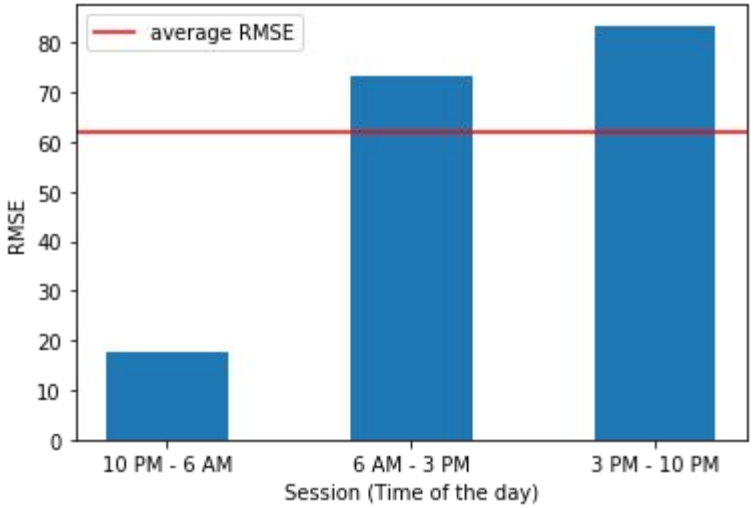
- Both feature set with or without PCA features perform equally well on our selected tree-based model.
- Although, Extra trees regressor overfits our training data with a r2 score of 1, it also gives, by far, the best performance on test set as well compared to other models, with a R2 score of 0.645.
- The test RMSE is also comparatively too low compared to other models.
- Hyper parameter tuning of our model doesn't have any significant impact on our test results. We were able to improve the test R2 score results by less than 1%.

# Analysis of our model

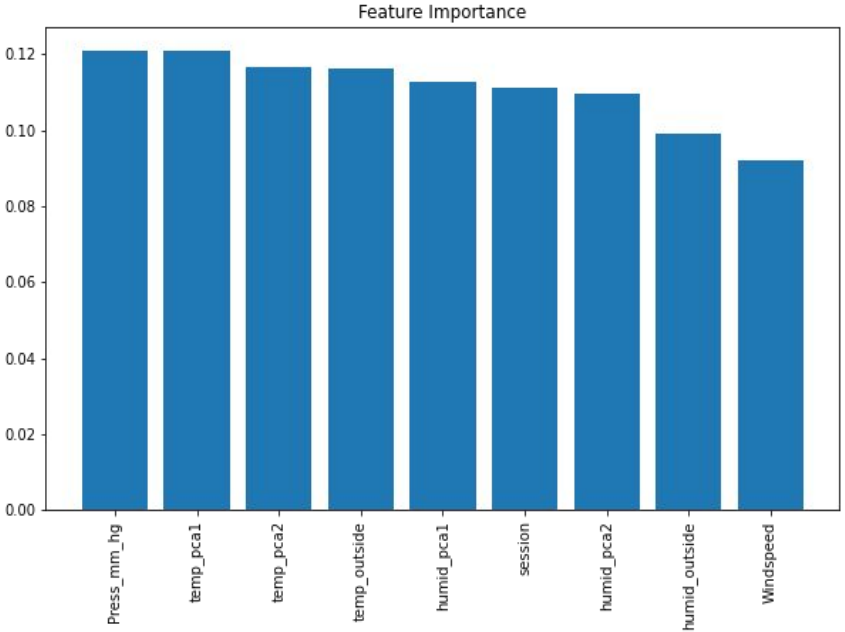


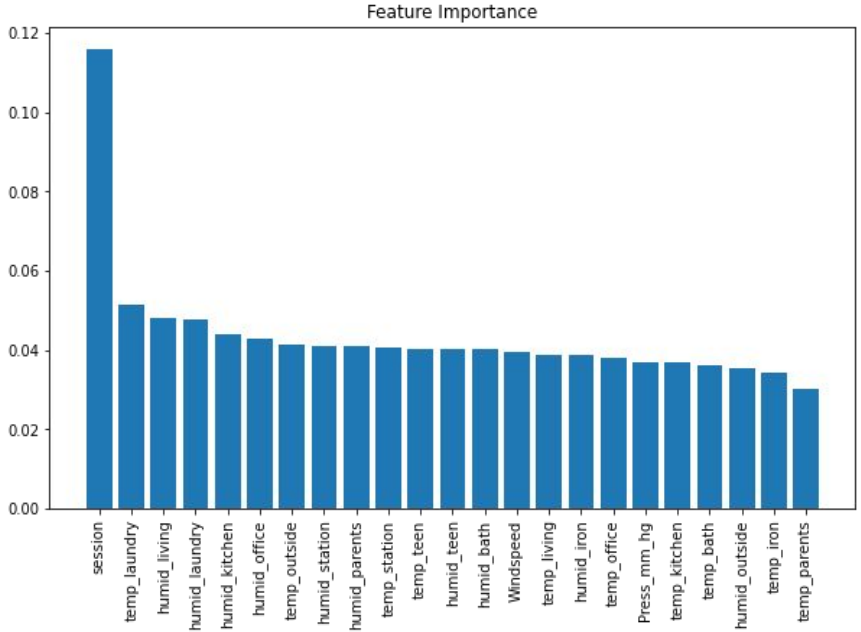
# Analysis of our model



# Feature importances



# Conclusion

1. The temperature/ humidity features showed little to no linear (pearson) correlation w.r.t target variable (<1%), although being highly linearly correlated within themselves.

2. The time zone of the day plays an important role in deciding power consumption of appliances.

3. The best Algorithm to use for this dataset is Extra Trees Regressor (tree based algorithm)

4. PCA helped us to reduce our feature set dimension considerably without affecting performance of our models significantly.

5. The untuned model was able to explain 64.5% of variance (R2 score = 0.645) on test set, while the tuned model was able to explain 64.9% of variance (R2 score = 0.649) on test set which is a tiny improvement of < 1 %

6. The least RMSE score on test data set is found to be around 0.6 by Extra trees regressor model, which is considerably good compared to other models.

7. Tree based models are by far the best model while dealing with data set that has most of its features having no linear correlation with target variable. For similar reasons, linear models such as linear regression, Ridge and Lasso perform the worst.


thank you!