TEAM - 10

INDEX

INTRODUCTION

Electrifying a nation

With 31 million Indian homes still without access to electricity, there is growing worldwide attention to extending electricity access to a much larger population. In order to address the expanding demand for a more sustainable and efficient energy source, automation technology (especially data analytics) is at the heart of a major digital transformation of the power grid, and helps by gathering data that allows us to make more informed decisions about energy use. Predicting electricity consumption can expand the scope in the following 4 major frontiers for an electricity generation company:

- **Energy Efficiency:** The forecasting of the electricity consumption at a portfolio level helps in estimating the expected amount of resources required and hence, the optimum utilization of the same.
- **User Satisfaction:** With an approximate estimation of the consumption of the user we can have a customized power supply for the user so that one does not encounter issues like short-circuiting.
- **Demand-side management:** The forecasting of demand gives us an approximate idea of the demand of electricity of either a user or group of people in a society and hence, helps in managing and catering to the demands of people.
- **Monthly bill in check:** Considering the user will have an idea about their expected consumption on a particular day. One can keep the bill in check by making amendments in the usage as and when required.

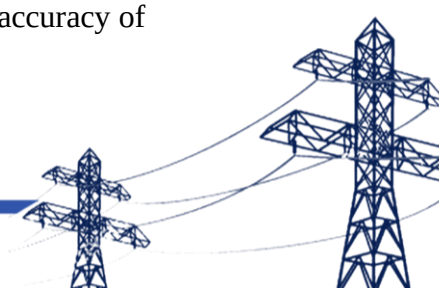
OBJECTIVES

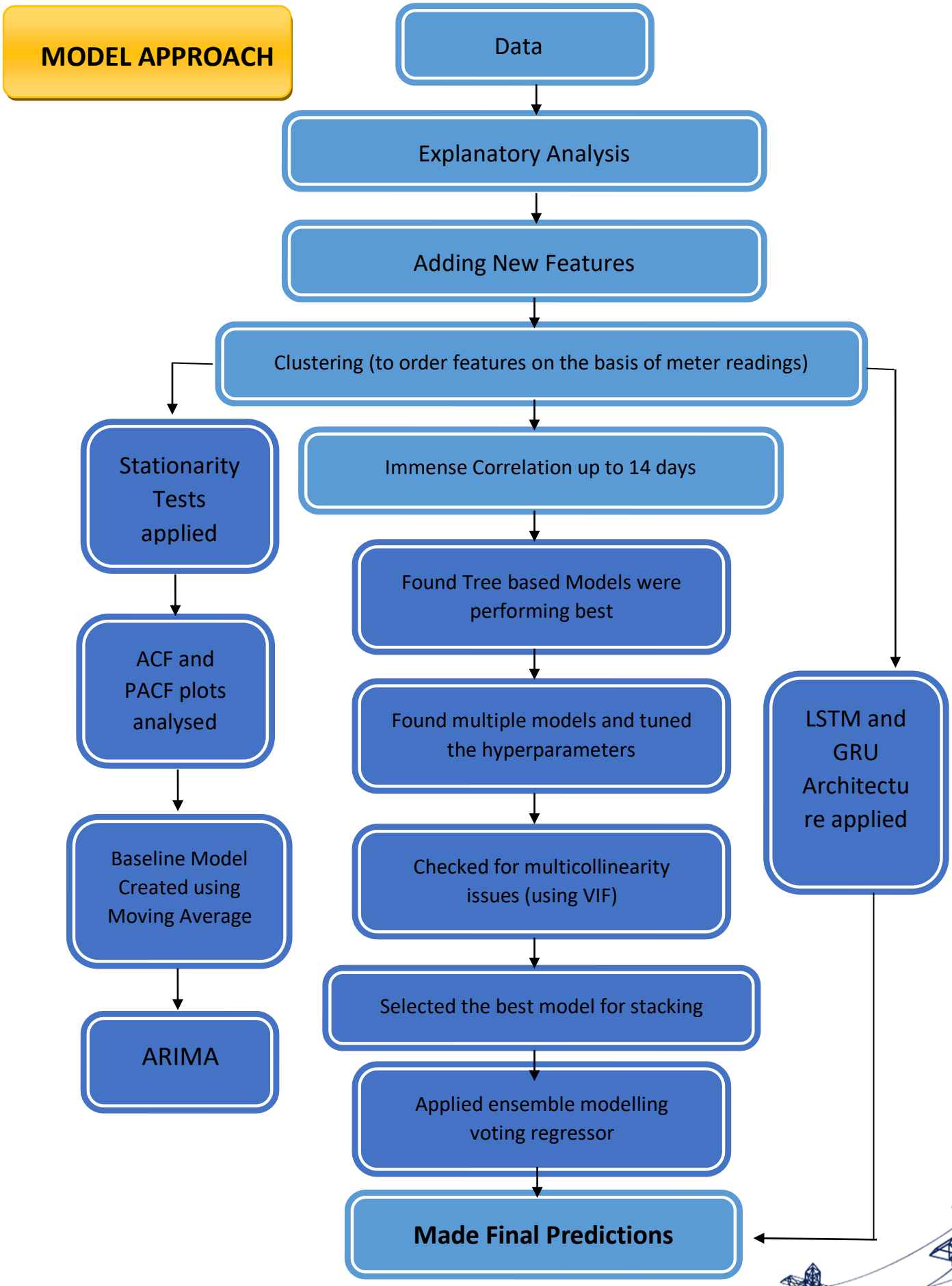
General Objective - Problem Statement

- Create a model to forecast the electricity demand for 5 corporate buildings in Delhi NCR.

Specific Objectives

- To uncover unique insights hidden in the data, about electricity demands in corporate buildings across the year in order to prevent wastage of electricity.
- To aggregate features to create relevant features and to find out which explanatory variables have a significant effect on electricity demand.
- To train and test different models and choose the best model based on accuracy of predictions.




MODEL APPROACH
Data
Explanatory Analysis
Adding New Features
Clustering (to order features on the basis of meter readings)
Stationarity Tests applied
ACF and PACF plots analysed
Baseline Model Created using Moving Average
ARIMA
Immense Correlation up to 14 days
Found Tree based Models were performing best
Found multiple models and tuned the hyperparameters
Checked for multicollinearity issues (using VIF)
Selected the best model for stacking
Applied ensemble modelling voting regressor
LSTM and GRU Architectu re applied
Made Final Predictions

DATA DESCRIPTION

The temporal data of electricity consumption for five buildings is measured across three meters from **April to December 2017** is given.

Timestamp

Meter readings are taken every 15 min, thus, 26,400 timestamps for each building.

Main_Meter

Continuous variable, showing the readings of the master meter installed in all 5 buildings.

Sub_Meter

Continuous variable, showing the readings of the sub meter 1 and 2 installed in all 5 buildings.

Building_no

Categorical variable, indicating building number.

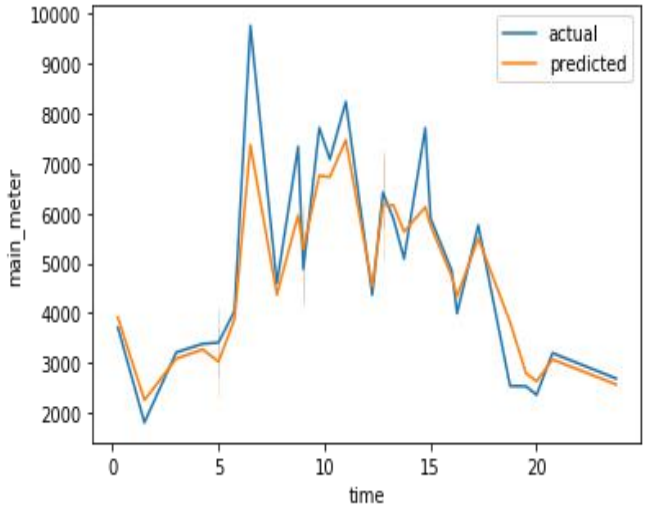
DATA PREPROCESSING

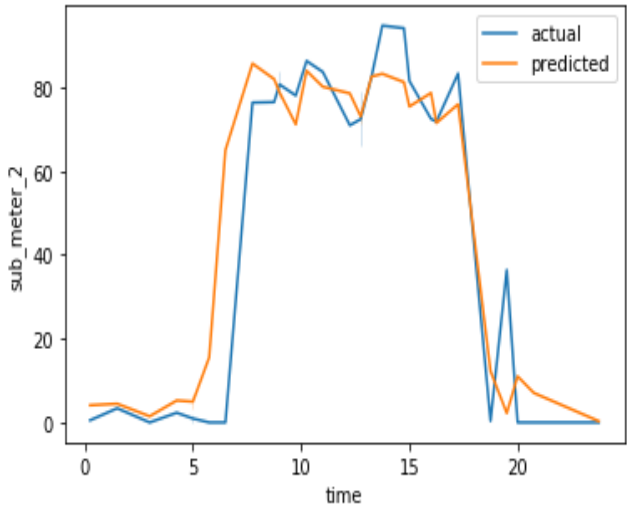
Feature Engineering

We created the following new features:

- **Holidays:** From the graphs plotted we find that the meter readings are significantly low on Sundays which means these days are holidays. We also consider that the Corporate Sector will be having holidays on some of the National Holidays.
- **Weekday:** From the graphs plotted, we find that the meter readings differ significantly over the week. This categorical variable indicates the weekday to which a timestamp belongs.
- **Seasons:** To expand the scope of our prediction to the months whose data was not provided, we created features for the different seasons we encounter throughout the year. This feature is intended to relate different months by their weather conditions and hence, to the electricity consumption.

- **Lag:** We calculated two types of lags for all the three meters which are as follows:
    - The first type of lag was calculated by taking the mean of every 24 hours for the past 14 days in total. Hence 14 lags were created for each meter.
    - The second type of lag was created with lag1 (shifting data by 24 hours) to lag14 (shifting data by 14*24 hours) for all three meters.
- **Working hours:** From the graphs, we inferred that electric power consumption was higher during the time interval (6:00a.m.-7:00p.m.). This can be attributed to the working hours of the corporations.
- **Month_fft, week_fft, day_fft:** They help us to capture the pattern month, week and day wise respectively. The values of **Fast Fourier Extrapolation** of all the meter readings. This indicates the trend in the meter value readings and hence helps reduce the contribution of the noise in the data to the results.





NOTE: Submeter sum: We calculated the sum of sub-meter 1 and sub-meter2 readings. The correlation of this variable with the main-meter reading is significant (around 0.7 for all buildings). This is also intuitive, because main_meter mostly measures the cumulative meter readings of all the sub meters (there are other sub meters as well in the corporate buildings).

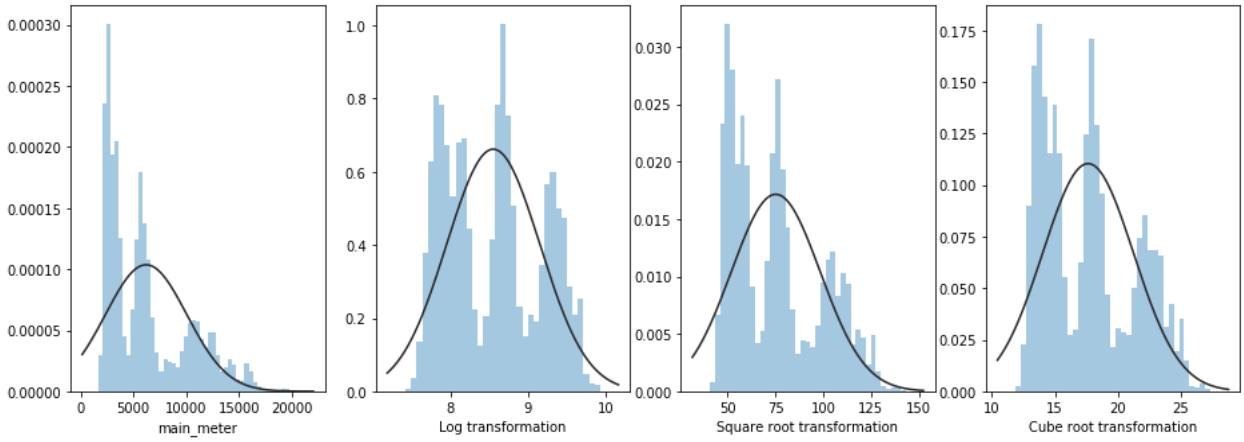
Feature Analysis

CHECKING STATIONARITY

As the given data was a time series data, we checked for stationarity, so that we can apply moving average, ARIMA and exponential smoothing.
Hence, we used **Dickey-Fuller** and **KPSS** tests.
Dickey-Fuller Test's results state that the data is stationary whereas KPSS states that the test is not stationary this implies that the data is **difference stationary**, later found the order of difference to be 1.

CHECKING SKEWNESS

Skewness is checked for better performance of the regression models applied in later stages.
Features like submeter 1 and submeter 2 appeared to be very skewed for different buildings. Hence, we used different transformations like logarithmic, square root, cube root (**Box cox)** and selected the one which gave the most standardized distribution.



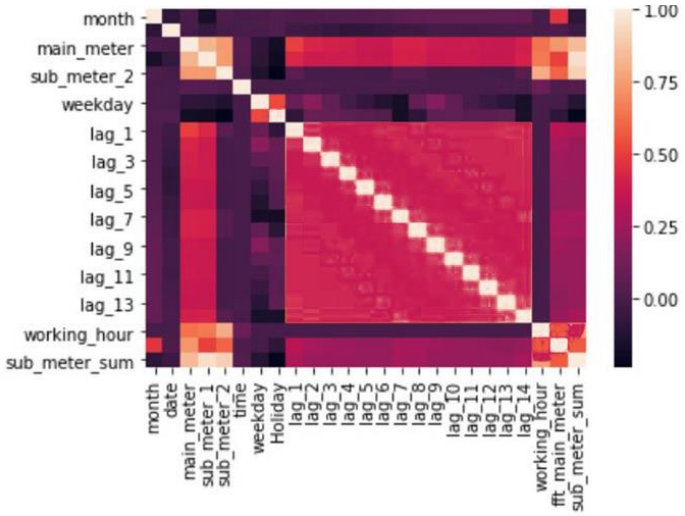
However, the skewness of main meter for Building 4 isn't apparent. Hence, **Welch Two Sample t-test** was performed. And p-value came out to be 0.25, So, null hypothesis was accepted (p-value>0.05) and it was concluded that it is not skewed.

Feature Selection

FOR CONTINUOUS VARIABLES

Correlation matrix and heat maps were used to find the significant features.

Heat Map for Building 1



FOR CATEGORICAL VARIABLES

The significant features are selected by first using Data Visualization Techniques like histograms, cat plots and box plots. After visualizing the graphical representation, we used hypothesis testing to get concrete proof for the same. Following tests were performed:

- **Two sample T-test** for the features holiday and working hours:

Building_1	**P-value**	
	Holiday	Working Hour
Main_meter	0.0	0.0
Sub_meter_1	4.199017979958388e-113	0.0
Sub_meter_2	0.0	0.0

We performed T-test for meter readings of other buildings also and p-values were coming out to be less than 5% which shows the feature significance. Since, the **p-value < alpha (significance value = 0.05)**, the Null Hypothesis (mean of the 2 samples are same) is rejected and features Holiday and working hours were selected.

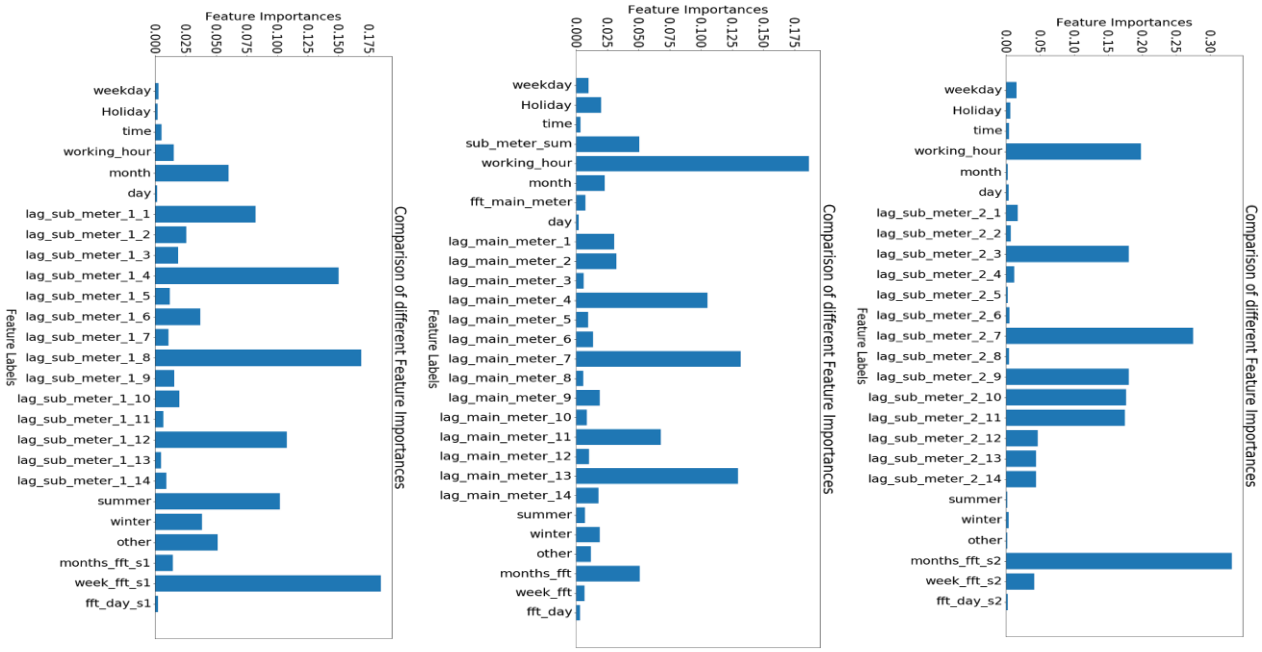
- **ANOVA test** for weekdays as it is a **multicategorical variable** over main_meter, sub_meter1 and sub_meter_2 in case of weekdays.

Building 1	**P-value**
Main_Meter	3.24258152181906e-260
Sub_Meter_1	1.8055479108462347e-87
Sub_Meter_2	0.0

As you can see, for weekdays also, the p value is coming out to be less than significant value, so we selected it as a feature.

Feature Importance

We used **Extra Tree Regressor** to assess the feature importance for both categorical and continuous variables.

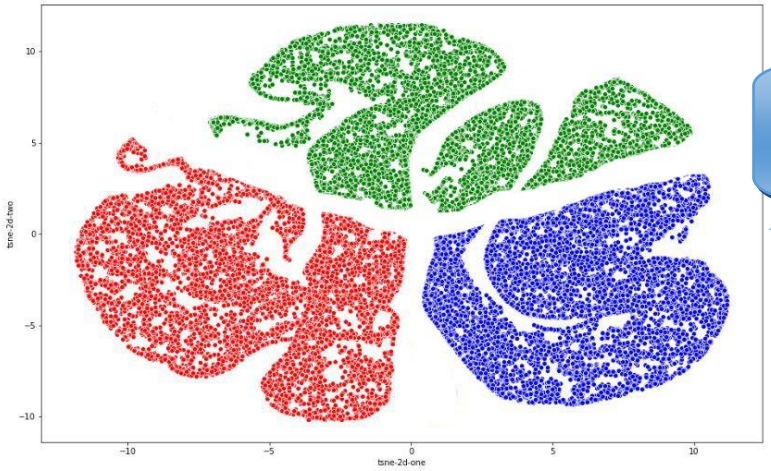


BUILDING 1: Sub_meter_1 Main_meter Sub_meter_2

CLUSTERING

We clustered the entire data into 3 clusters to identify the discrete months, day of the week and date of the month belonging to homogeneous groups. We try to visualize it using **t-SNE plots** which convert the high dimensional data (25) into two dimensions suitable for human observations.

We find the points in Cluster-1(Red) to have the highest electricity consumption. Cluster-2 (Blue) has medium electricity consumption whereas Cluster-3(Green) has the lowest electricity consumption of all.

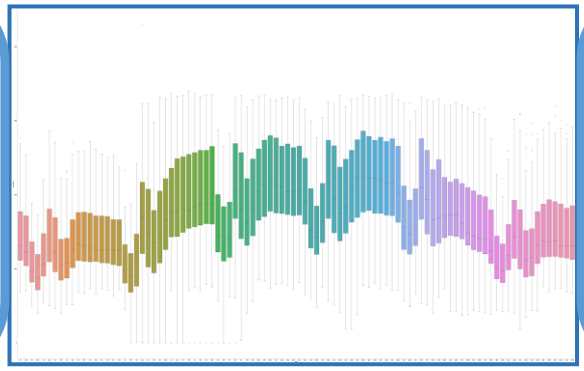


Silhouette Score:

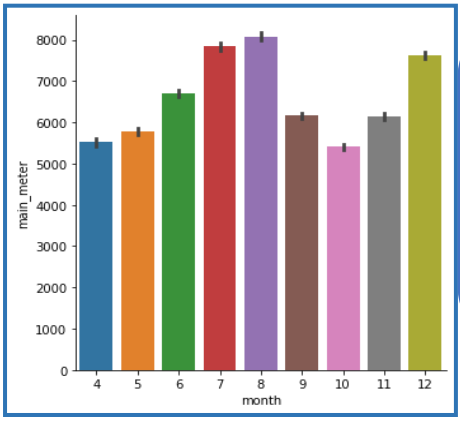
0.61

INSIGHTS

In **Building 4**, the pattern is seen in **Submeter 1 and not Submeter 2** and this could be attributed to an exchange in the appliances these meters measure.

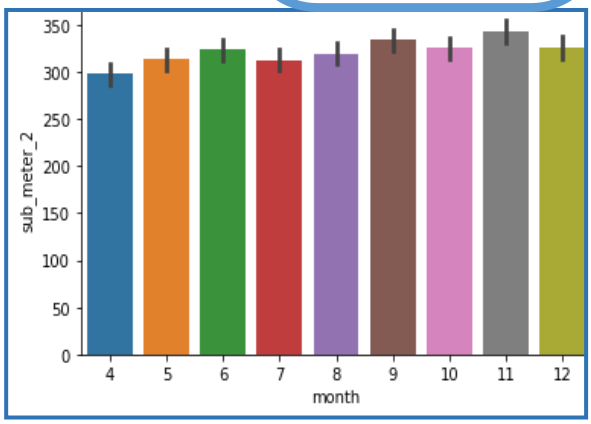


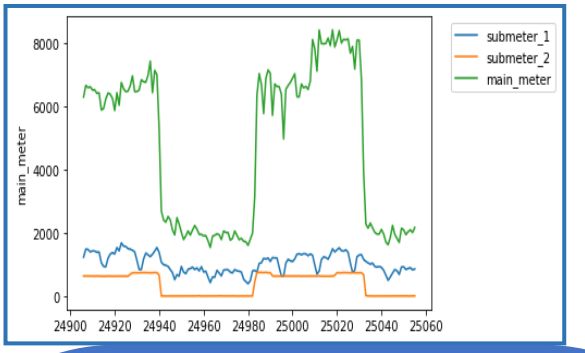
For all buildings except 4, Submeter 1 readings show a **local minima after every 4 hours**. This drop **lasts for 30 minutes**. This indicates that there might be workplace breaks every 4 hours in 30-minute slots.



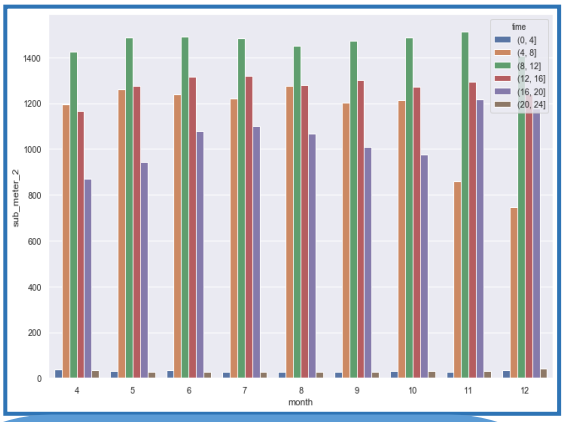
The main-meter readings were found to be less for the **months of April, September and October**. Even for the month of May, the main-meter reading was not as high as expected given the summer heat.

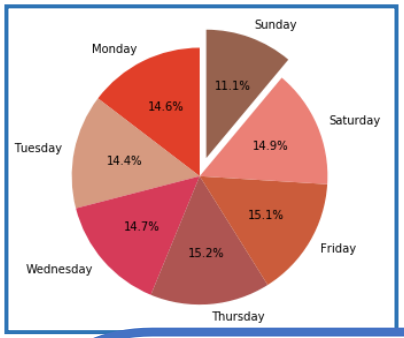
The **sub-meter 2** reading was found to be **nearly constant** for all the months across all the buildings.





Main meter readings are quite high. **Submeter 2** readings are significantly **lower than** the main meter and **Submeter 1** readings. A possible explanation for this is that the submeter 2 is used to monitor energy usage for individual departments or pieces of equipment to account for their actual energy usage.





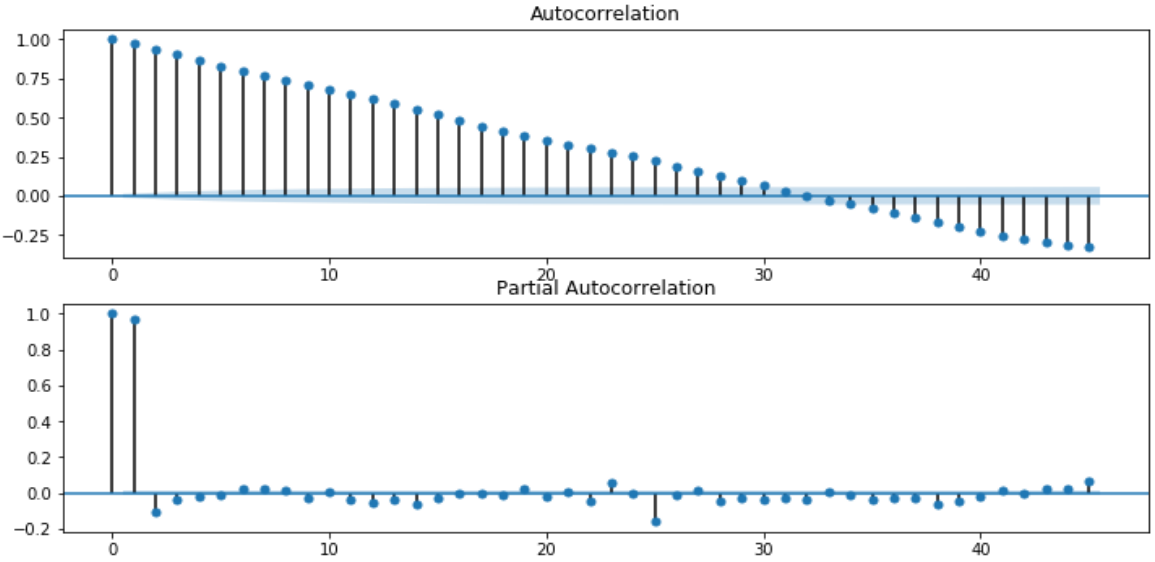
In all buildings **except Building 4**, the **Submeter 2 readings are significantly low** during non-working hours and show a **sudden increase few hours before and after the working hr** (6AM to 7PM). This implies submeter 2 measures an appliance or department that is switched on a few hours prior to the arrival of employees in the morning and is switched off a few hours after the departure of the employees after work.

The electricity consumption has a **noticeable dip on Sundays**. This implies that Sunday is a day off for most employees of these corporate buildings, but Saturday isn't.

MODEL TRAINING

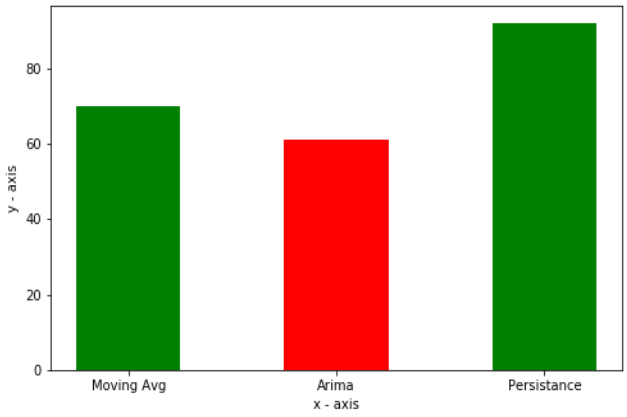
Baseline models

We initially created baseline models using **Persistence Algorithm, Moving Average and ARIMA.**
For proper values of **p, d, q** parameters of the **ARIMA** model we analysed the **ACF** and **PACF** plots which helped us to conclude that the ARIMA model **(2,1,0)** is the most optimal for all the buildings and all the meter readings. Since the data does not have a visible trend or seasonality it does not appear to be an ideal candidate for Holt's Winter Models.



(For Building 1)

Evaluation Metrics for Baseline Models

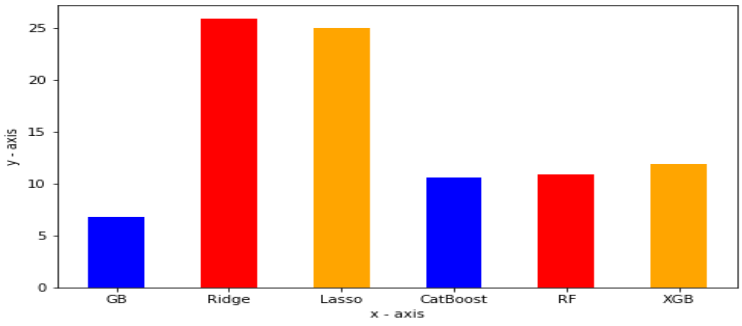


Regression models

We applied several regression models among which **Gradient Boosting**, **Random Forest** and **Xgboost** gave very good results. The following models were tested:

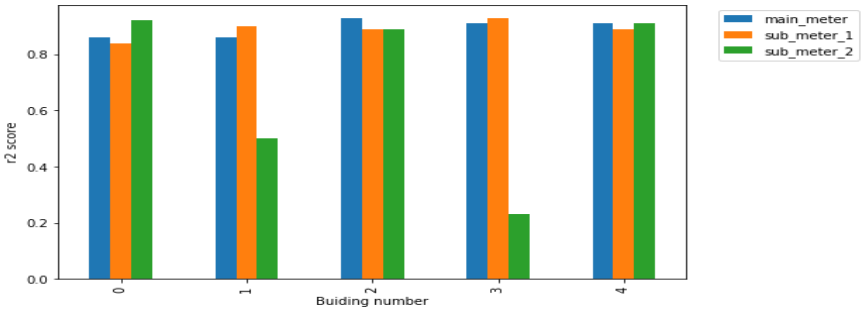
- Lasso Regression
- Ridge Regression
- Random forest Regression
- Elastic net regression
- Support vector regression
- XGBoost regressor
- Gradient Boosting regressor
- Catboost Regressor

NOTE: Considering, we are using lagged values as a feature in our regression models there could be an **issue of multicollinearity**. Hence, we checked for multicollinearity between the variables using **VIF (Variance Inflation Factor)** and found that the VIF values for all buildings except 4 was coming out to be **less than 10** (**less than 5** for some meters). So, the lag values were removed for building 4 and used for the rest of the buildings.



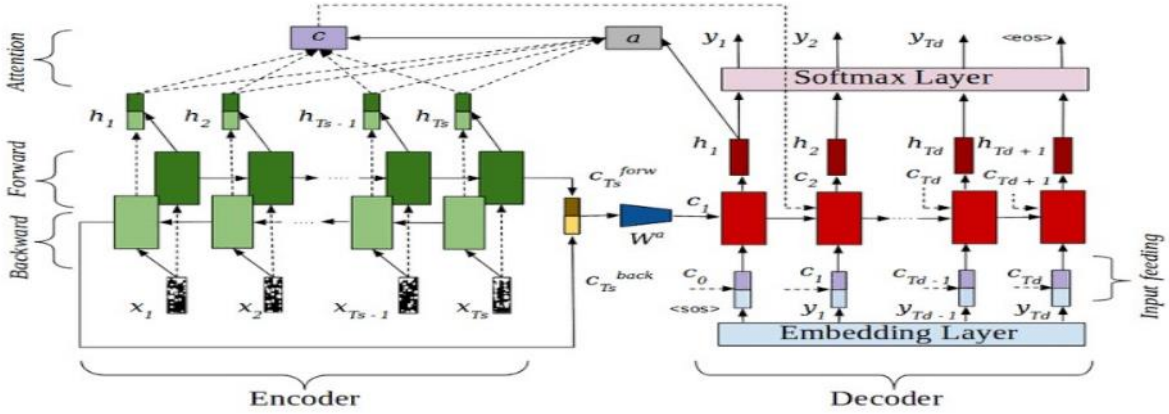
Evaluation Metrics for various Models

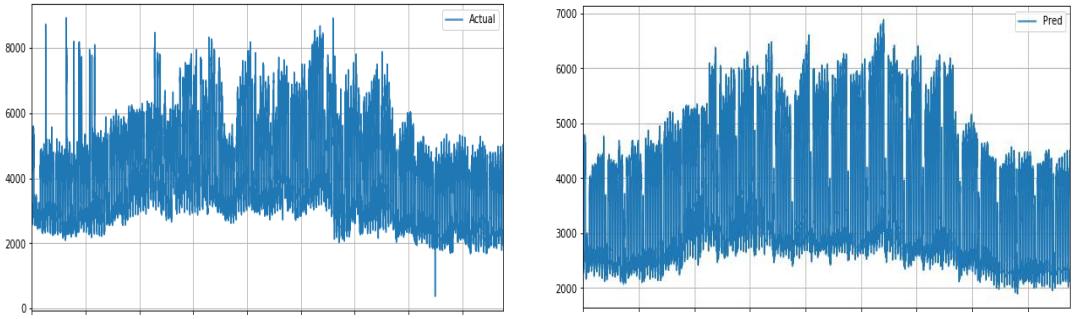
R2 Score for Gradient Boosting



LSTM

Once we have created the vectors for different time-stamps we can pass them through an **Encoder-Decoder Long Short-Term Memory with Attention Layer**. We use LSTM to capture the pattern of electricity consumption by the customers. The reason behind using **Attention layer** is to **focus more on certain parts of the input when predicting a certain part of the output sequence**. After performing the necessary feature creation, **one hot encoding** of categorical variables and scaling of the numerical input features (lags), an LSTM model was used to predict the readings of sub meter 1 and sub meter 2 separately followed by prediction of the main meter readings using the lags, fft and the predicted readings of the two sub meters.
In the encoder phase we provide the **25-dimensional vector** as an input which encodes and outputs an encoded state. Whereas, in the decoder phase we provide the 24-dimensional vector (all features except the electricity consumption) as input and expect the output as the electricity consumption. So, using this architecture we provide the features for 14 days in the encoder phase and forecast the electricity consumption for next day keeping sufficient stride between 2 series of inputs. We can also have an auto-encoder layer along with decoder so that the encoded state is captured more efficiently.





(For main meter of Building1)

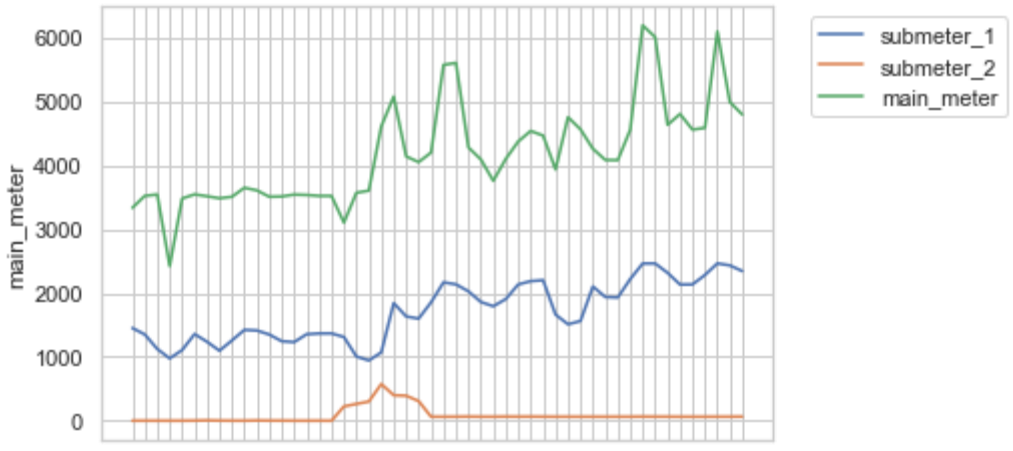
Note: According to the evaluation metrics, more weightage is given to correct prediction of electricity consumption on initial days of the month. Hence, we select model on the basis of lower weighted RMSE.

RESULTS

Model Name	Evaluation Metric	R2 Score (For Building 1)		
		Main_meter	Sub_Meter_1	Sub_Meter_2
Gradient Boosting	3.68	0.86	0.84	0.92
XGBoost	9.7	0.86	0.90	0.91
Random Forest	10.05	0.87	0.91	0.86

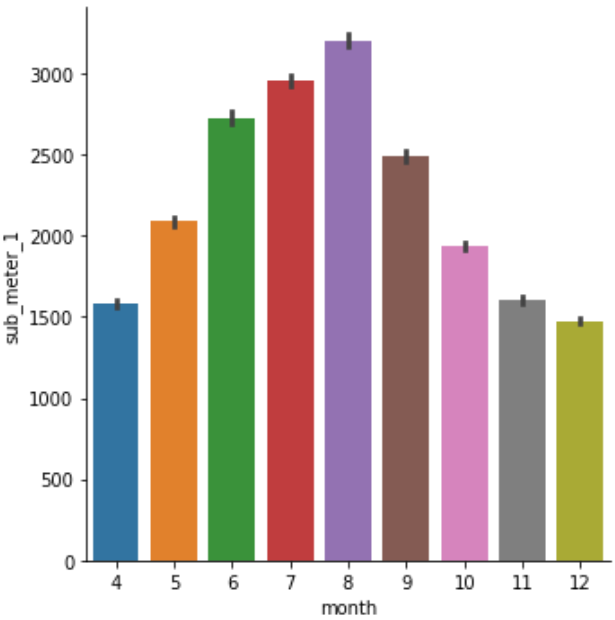
Model Name	Evaluation Metric	RMSE (For Building 1)		
		Main_meter	Sub_Meter_1	Sub_Meter_2
LSTM	16.64	3281	258	125.324

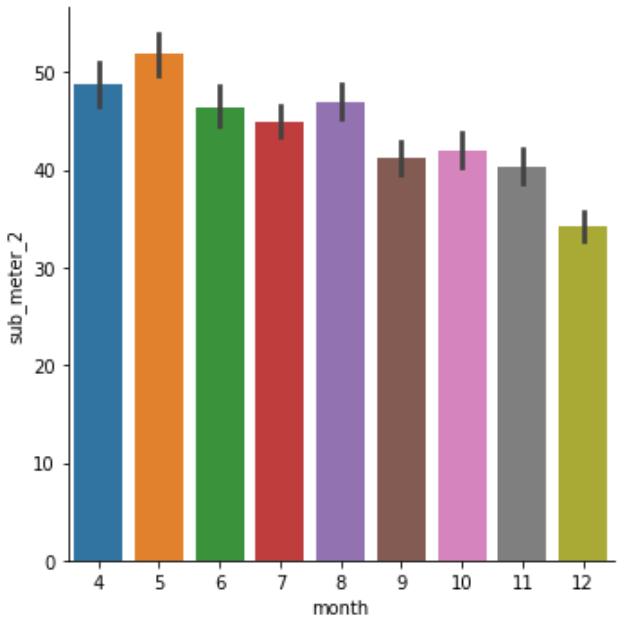
The plot for the predicted values of the given test set using **Gradient Boosting for the month of January**:

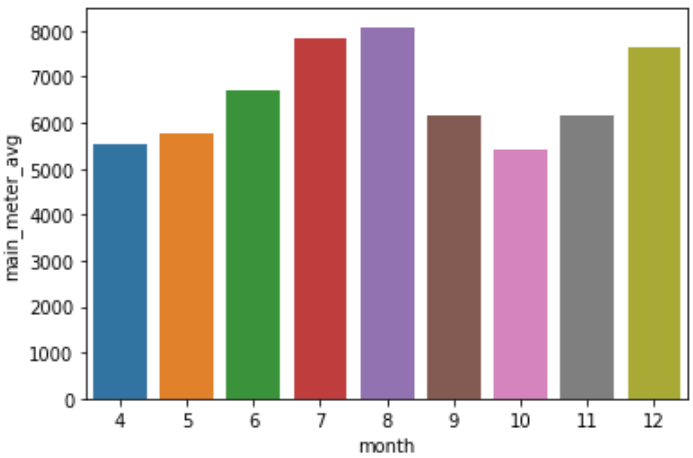


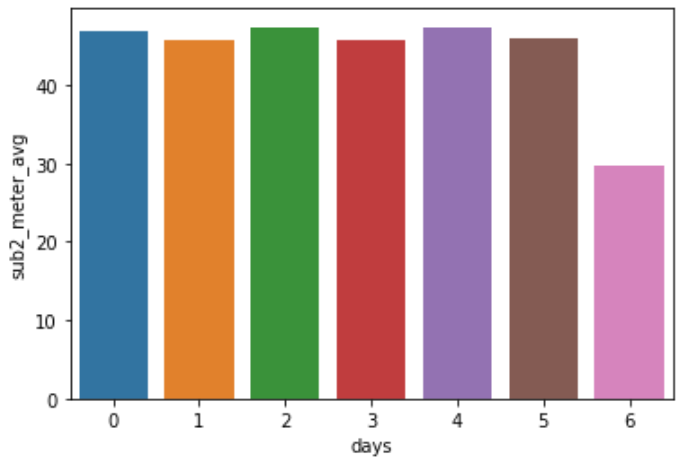
ANNEXURE

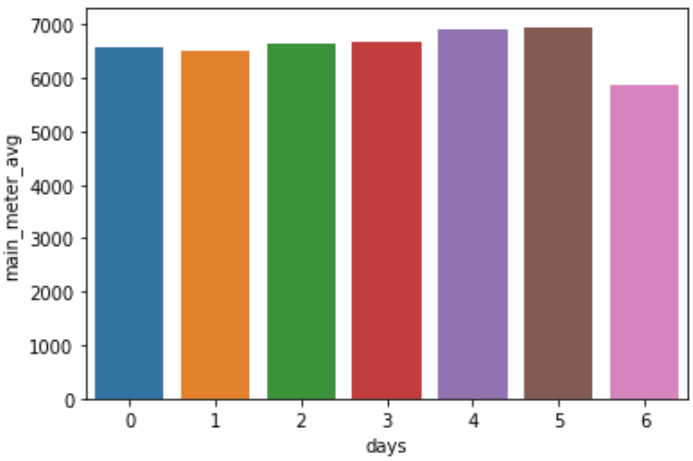
Building_1

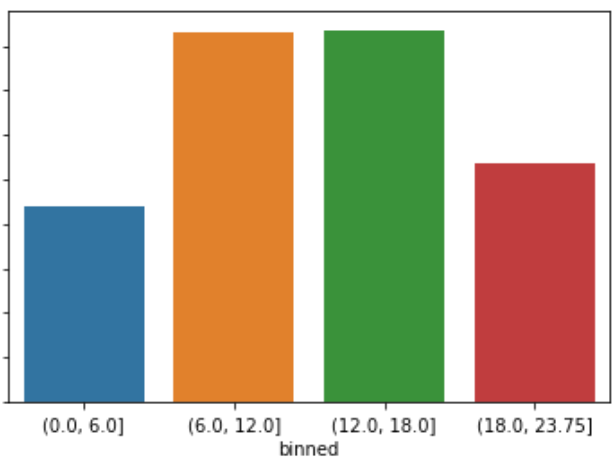

sub_meter_1
month


sub_meter_2
month

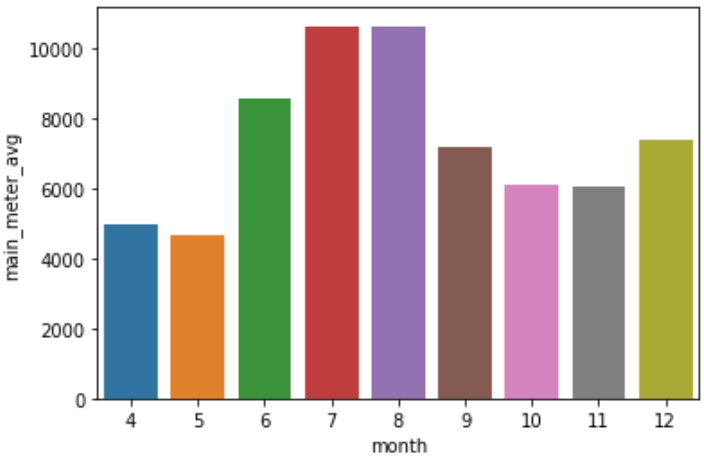

main_meter_avg
month

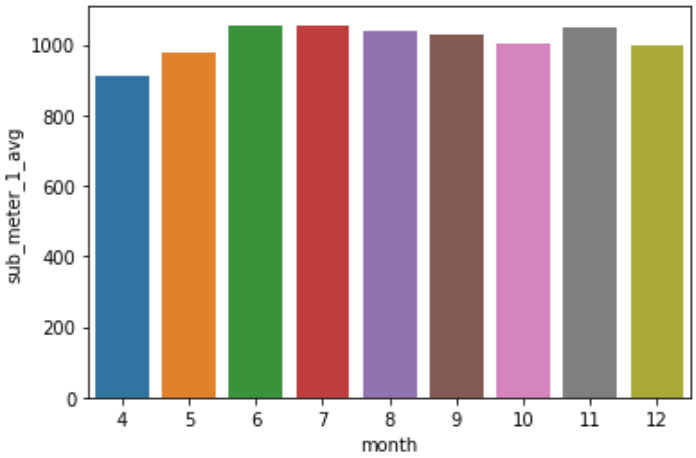

sub2_meter_avg
days


main_meter_avg
days

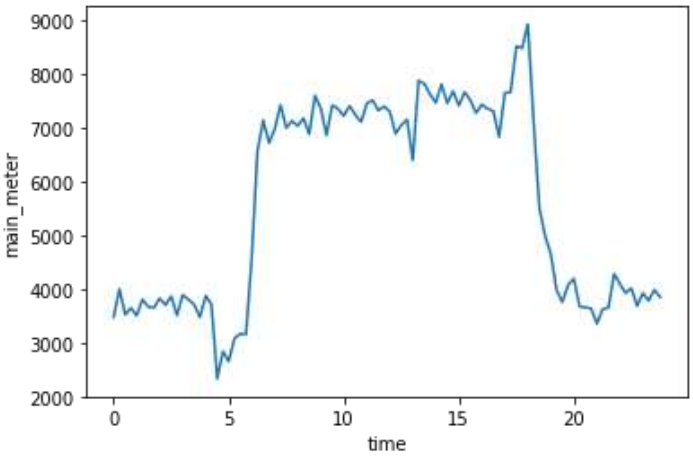

(0.0, 6.0]
(6.0, 12.0]
(12.0, 18.0]
(18.0, 23.75]
binned

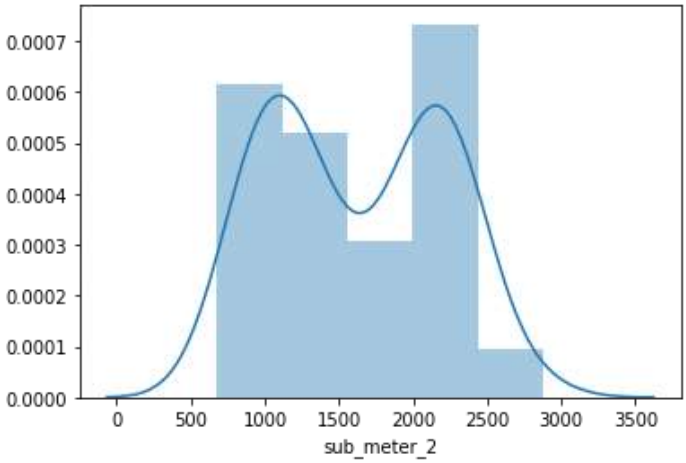
Building_4

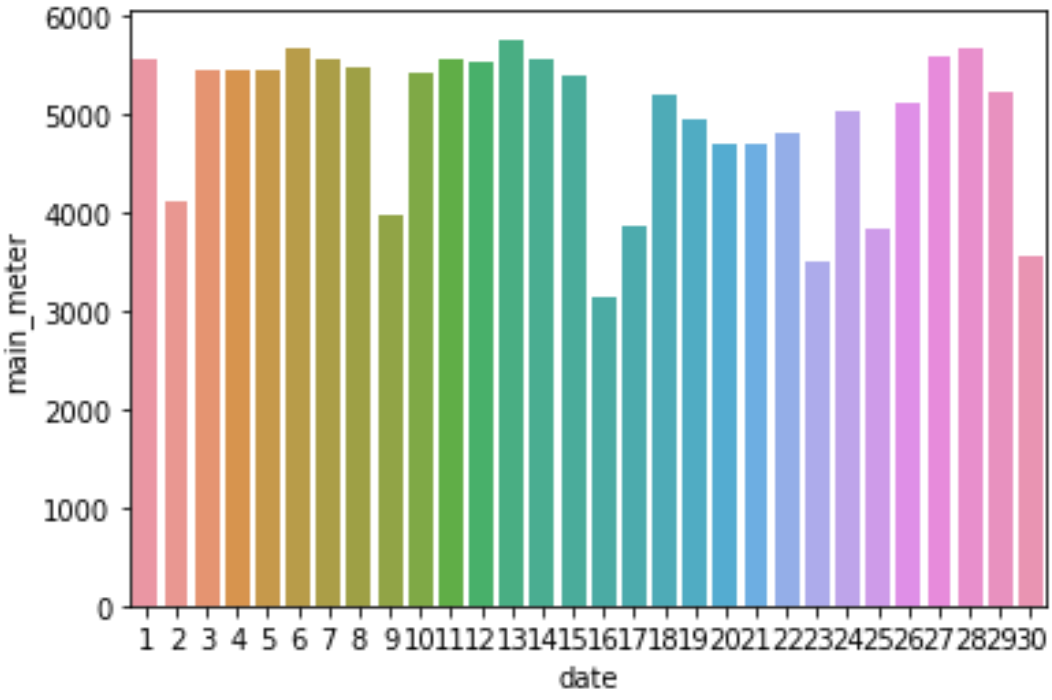


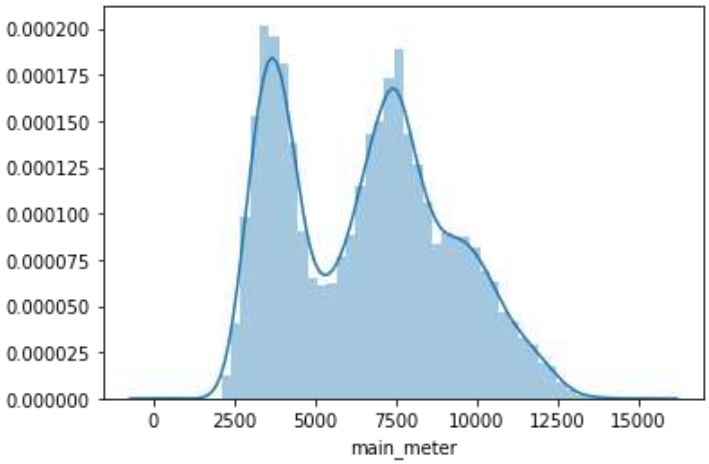


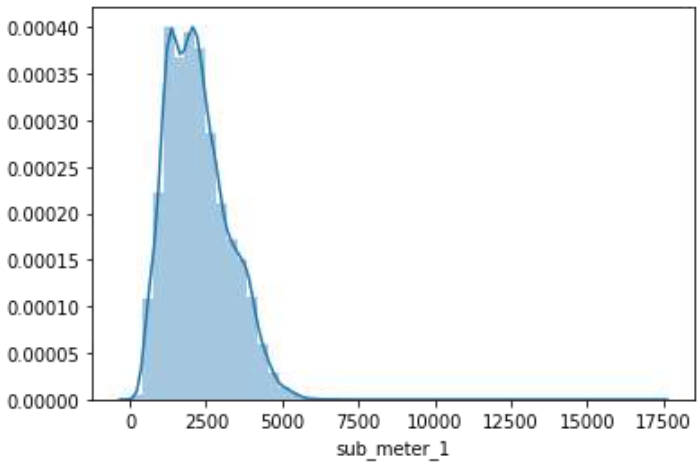
April 1 main meter readings

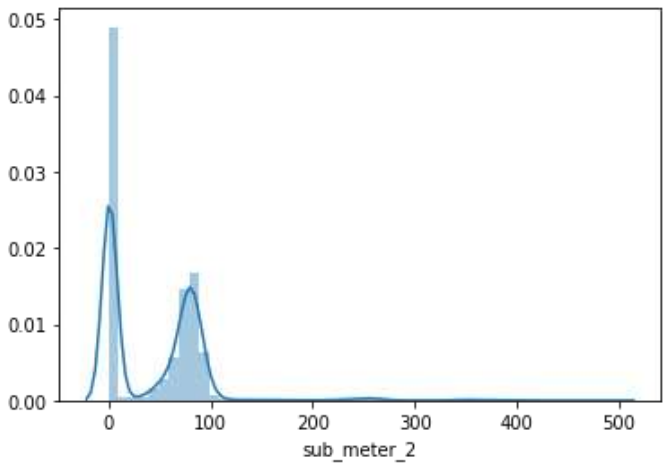


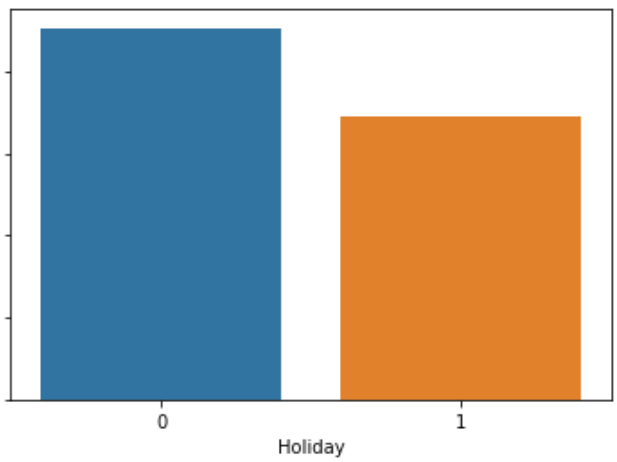


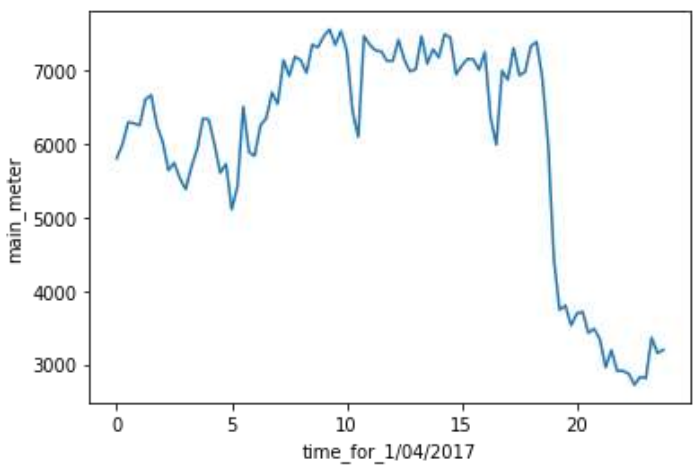


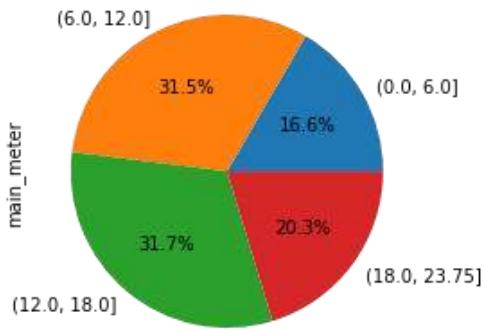

0.000200
0.000175
0.000150
0.000125
0.000100
0.000075
0.000050
0.000025
0.000000
0
2500
5000
7500
10000
12500
15000
main_meter


0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
2500
5000
7500
10000
12500
15000
17500
sub_meter_1

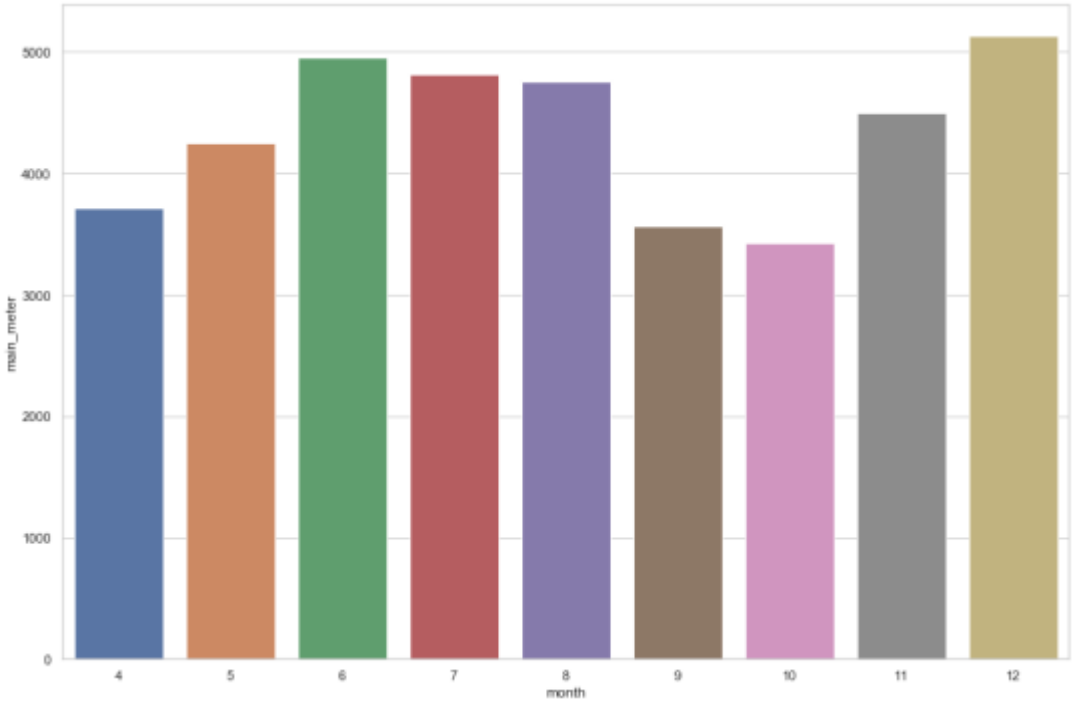

0.05
0.04
0.03
0.02
0.01
0.00
0
100
200
300
400
500
sub_meter_2

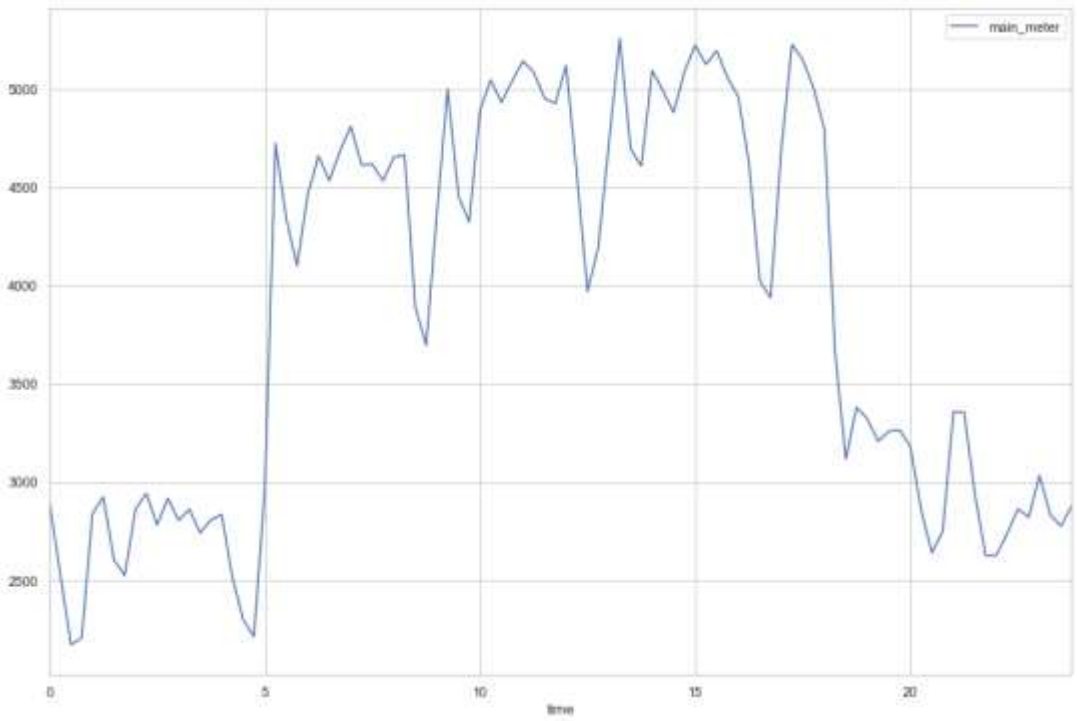

4000
3000
2000
1000
0
0
1
Holiday

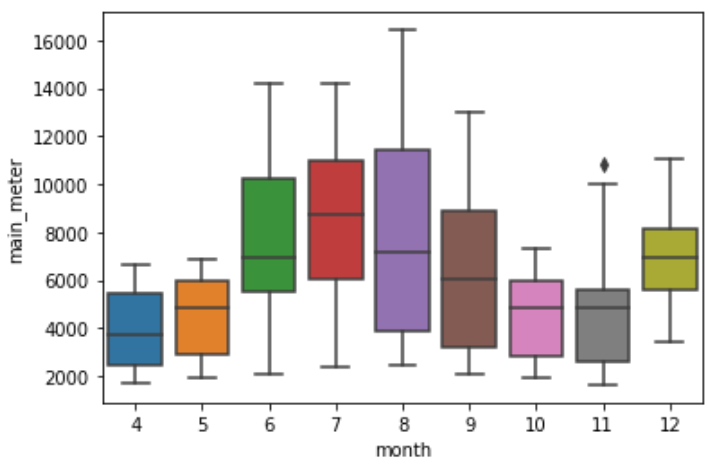

7000
6000
5000
4000
3000
main_meter
0
5
10
15
20
time_for_1/04/2017

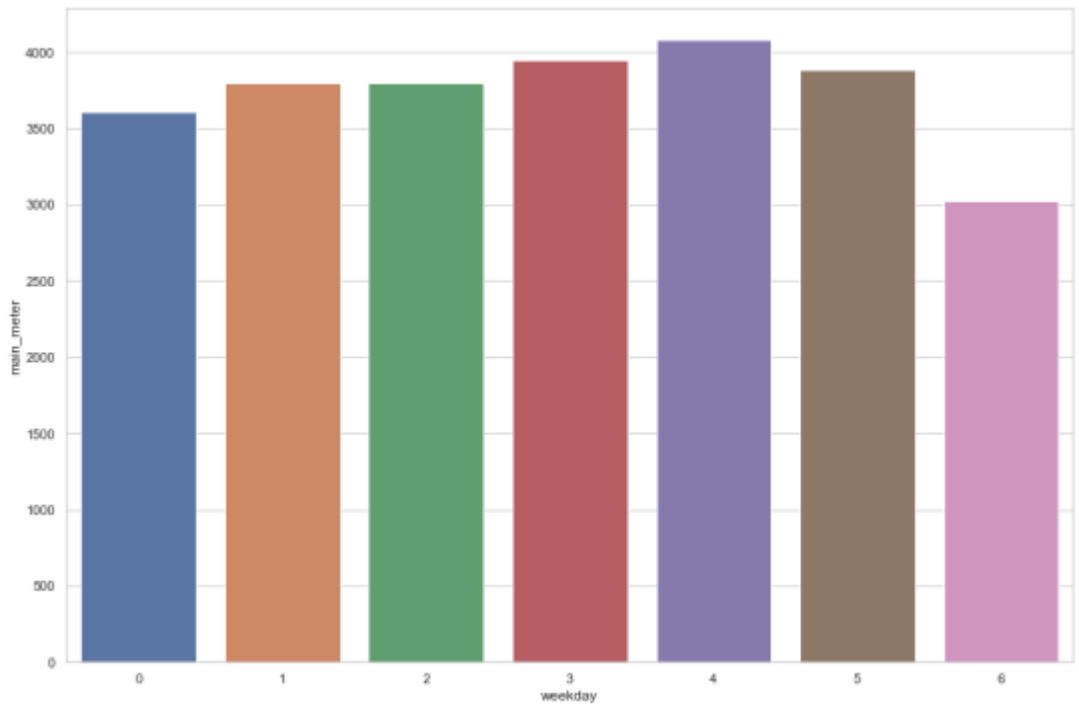

(6.0, 12.0]
31.5%
(0.0, 6.0]
16.6%
main_meter
31.7%
20.3%
(18.0, 23.75]
(12.0, 18.0]

Month wise plot for Building 1

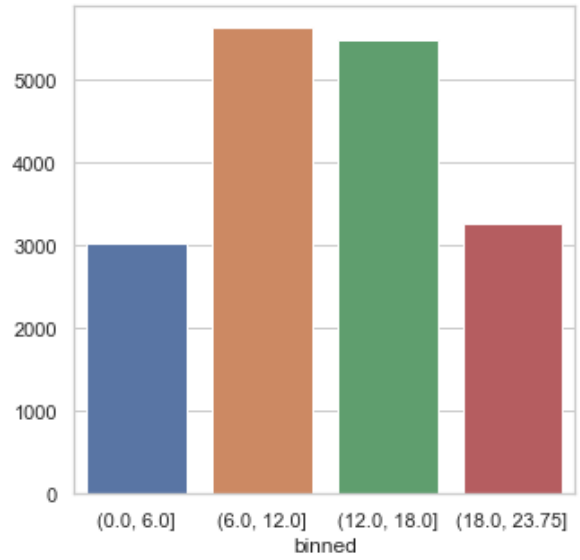


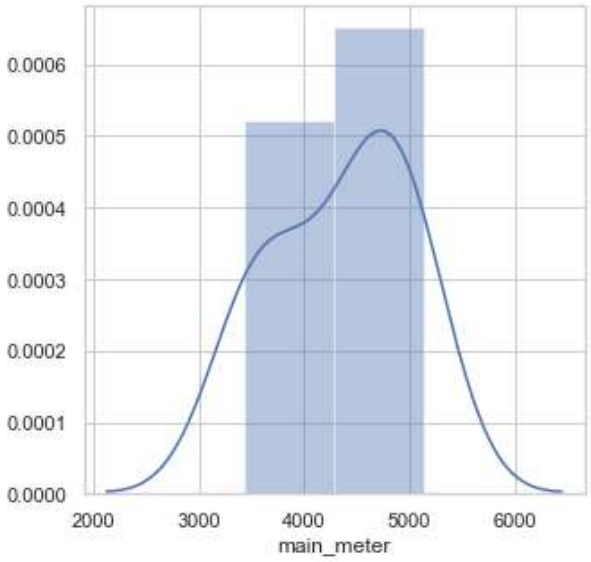




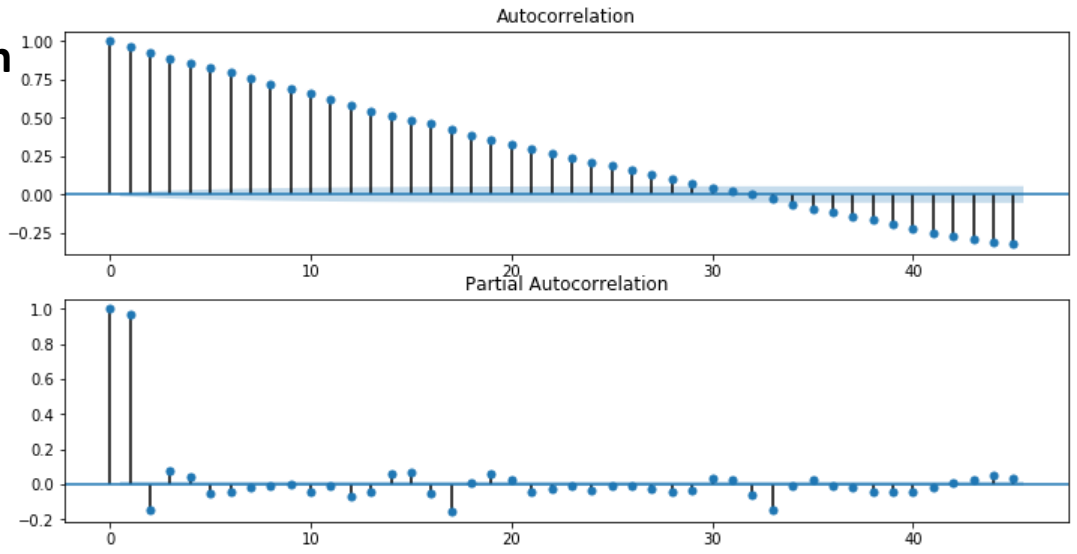


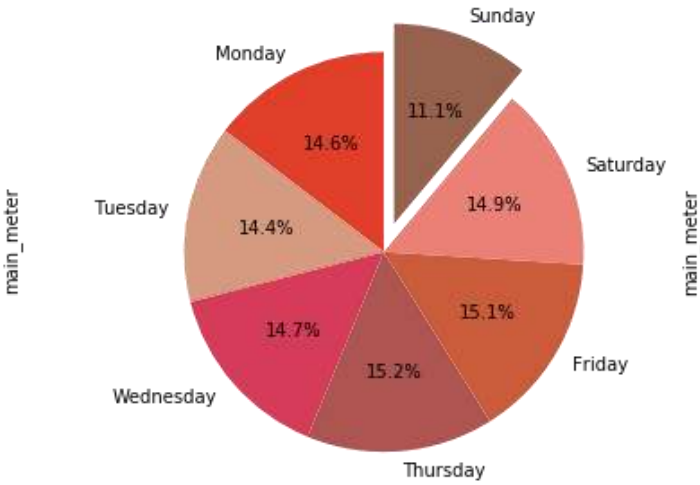
Building:1 Main meter readings Month:4 significant drop on sundays

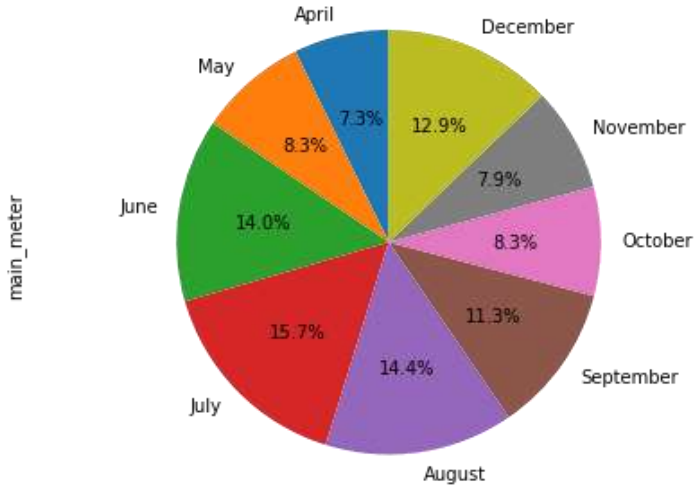


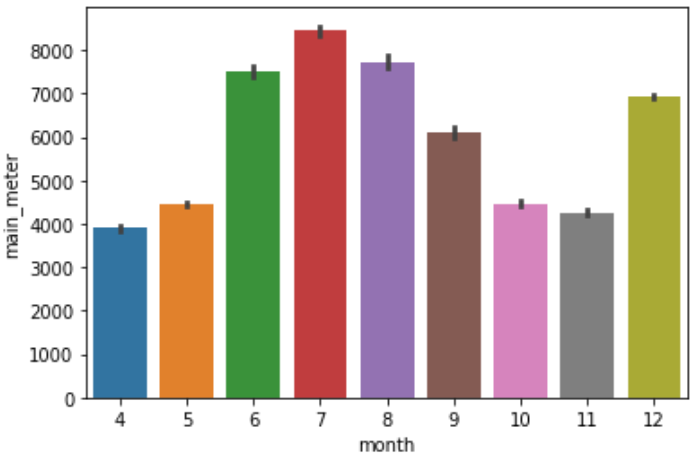


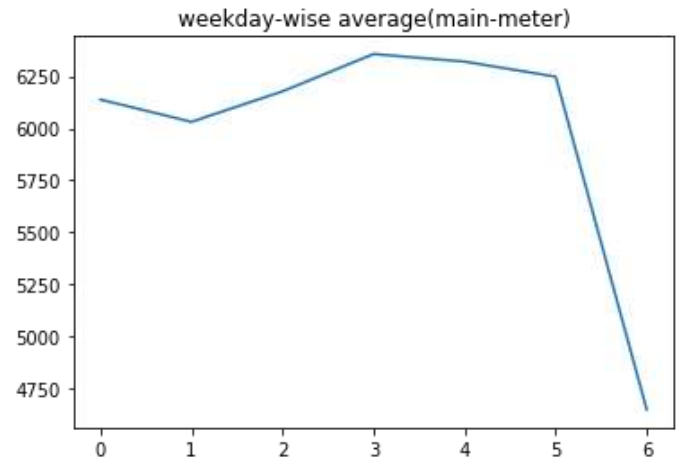
Building:1 6hrs interval max consumption during 6-12



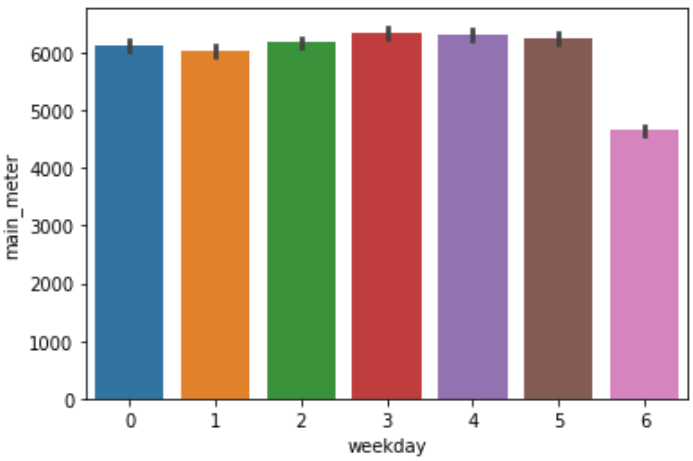


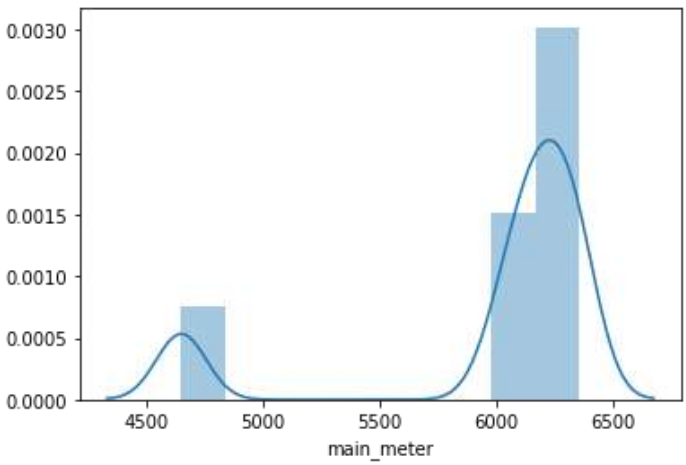




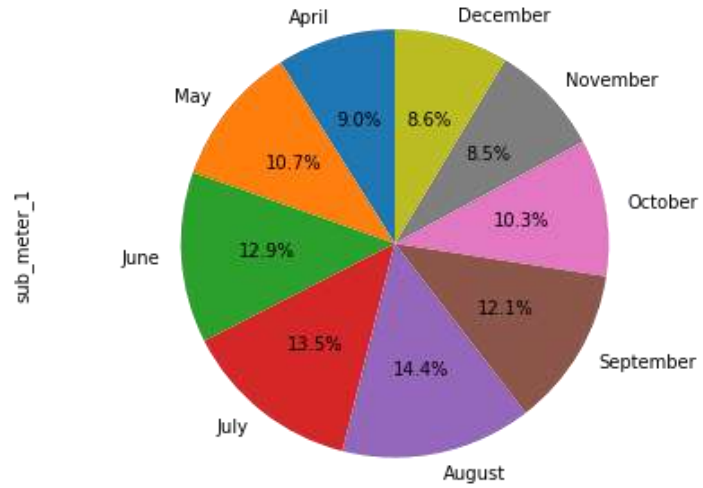


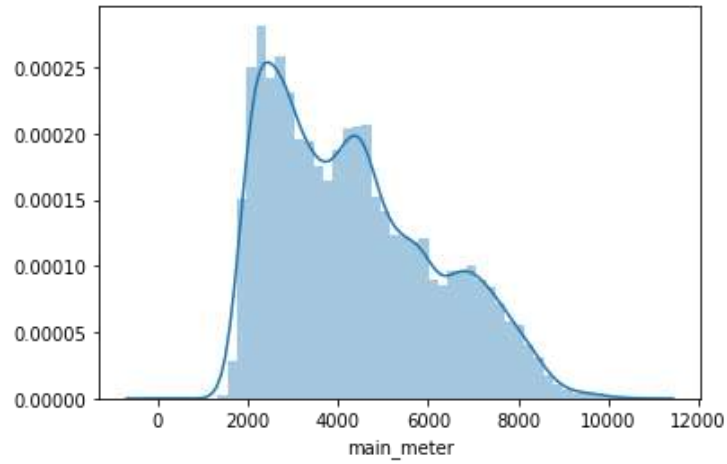
month-wise main-meter reading

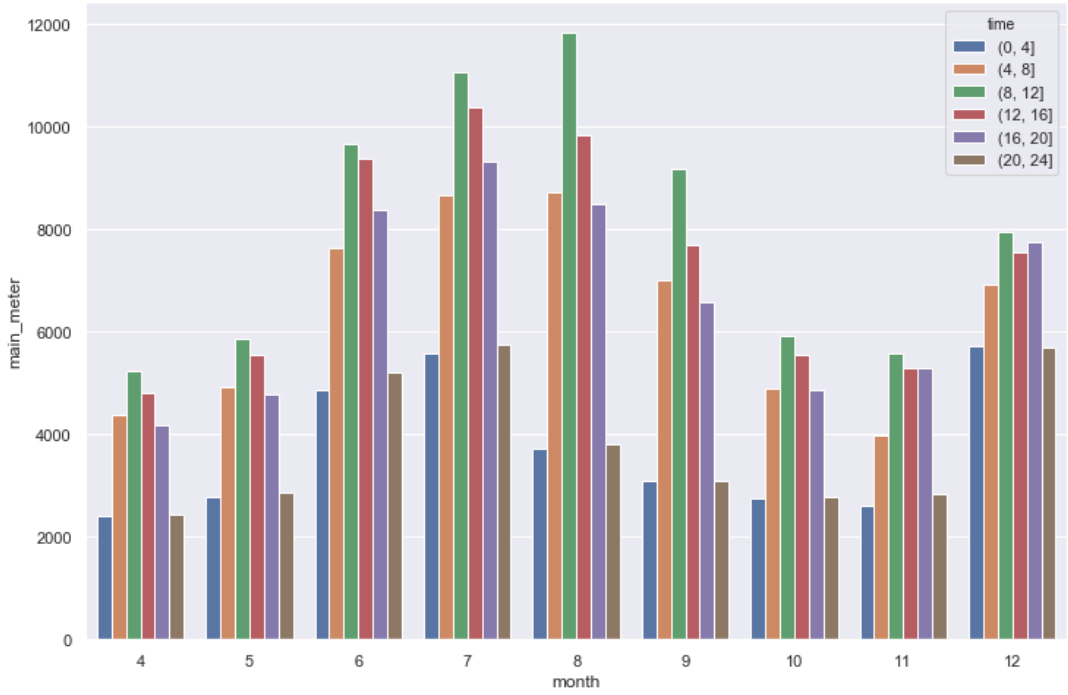


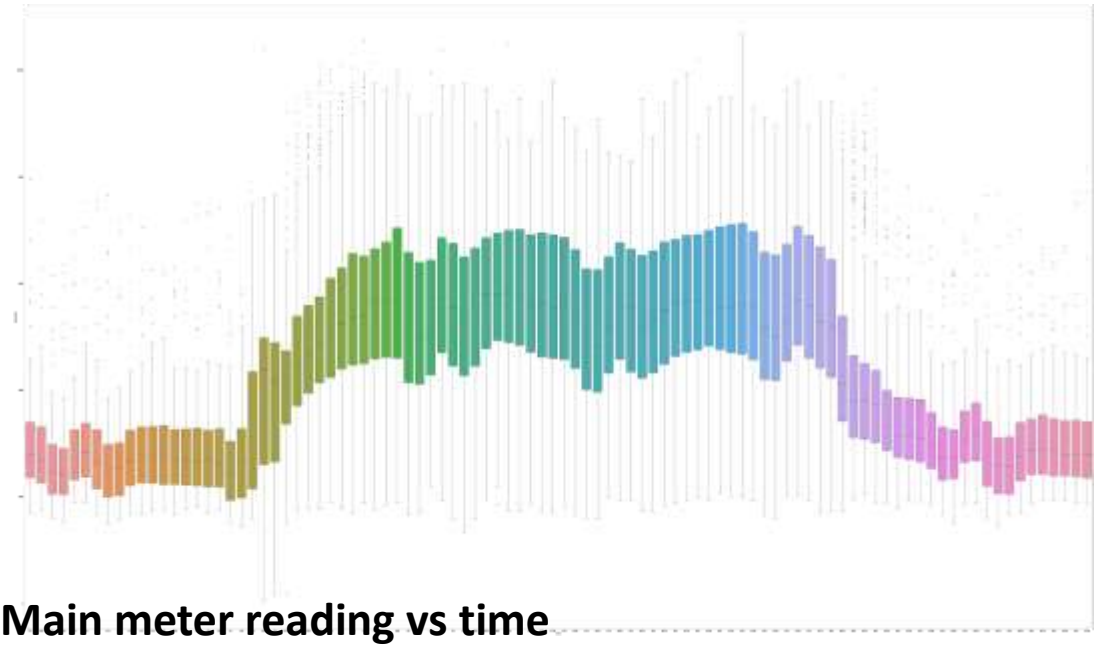


Distplot of weekday means









Main meter reading vs time