**Case Study 1 : Energy Forecasting for Boston City**

**INFO 7390 : Advances in Data Sciences/Architecture**

**Team 7**

**Chandni Sharma**

**Hari Panjwani**

**Sanath Shetty**

**Table Of Contents**

1. Problem Statement 4
2. Data Wrangling & Cleansing 15
3. Data Visualization Using Tableau 30
4. Multiple – Linear Regression 40
5. Forecast 50
6. References 60
7. **Problem Statement**

The aim of this project is to build a predictive model to forecast the energy consumption for the city of Boston. The raw data available currently has a record of power usage for every 15 minute interval in a day.

* 1. **Goal:**
* Analyze the existing energy consumption data for the city of Boston and develop an effective approach that can be used to monitor, plan and predict the power consumption for the future.
* Show trends in energy consumption based on various factors that are directly proportional to the power usage
  1. **Applications:**
* Recommend which account number or facility needs to be monitored and educate the citizens with suggestions for optimal usage of energy
* Make a note of consumption trends with respect to temperature, day of the week, month or even a particular day
* Recognizing a pattern in the usage to identify peak hour for daily energy consumption

1. **Data Wrangling & Cleansing**

As we are aware that most of the datasets available have a large number of records, variables etc. & there is a very high probability that the datasets would be **incomplete, noisy and inconsistent**. In order to make the data suitable for analytics we needed to pre-process the data to check for thee above inconsistencies. Before pre- processing the data we did a little research about it and discussed the data features that we needed to incorporate in our final table.

Blow is the summary of the raw data that we had received for this assignment.

|  |  |
| --- | --- |
| **Factors** | **Description** |
| kWh | kWh is aggregated hourly. Sum of 12 obs |
| month | 1-12 => Jan-Dec |
| day | 1-31 |
| year | Calendar Year |
| hour | 0-23 (Each rec corresponding to the hour) |
| DayOfWeek | 0-6 => Sun- Sat |
| Weekday | 1. Yes 2- No |
| Peakhour | 7am to 7pm ->1 ; 8pm to 7am -> 0 |

As we were planning to implement the prediction model in our analysis, we separated these columns by categorizing them into **“Must have”** & **“Keep for later”**

The columns that we included in our **“Must have”** are **kWk, month, hour** & **Peakhour**. The rest were categorized in the latter category.

As it is an obvious fact that power consumption is directly related to the daily temperature, we needed to acquire the temperature dataset for that particular year. We pulled the weather dataset from **WeatherData** package in R.

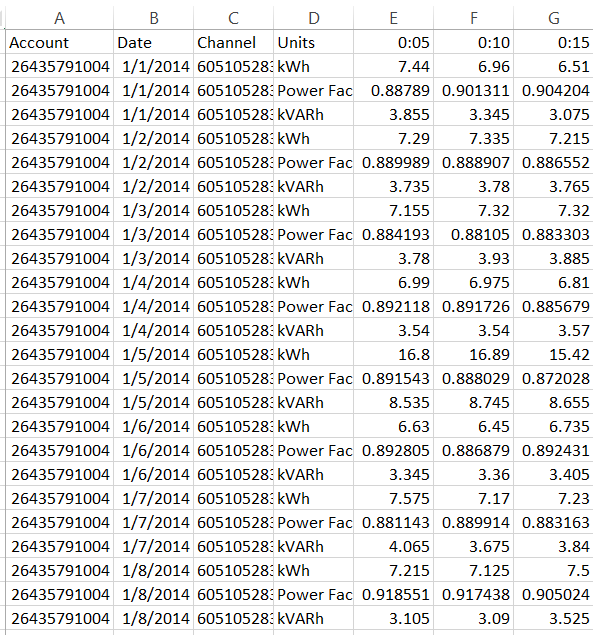
We identified two datasets, the first one being the rawdata.csv provided for the assignment and the second one being the newly acquired temperature dataset. **During the data cleansing procedure, we built functions to read csv files and clean them, transposing the columns, applying the melting function etc. in R for the two datasets.** While dealing with the datasets we encountered various problems with the data such as **“NA”** values, duplicates, **“0’s”** etc.

Our next step was to clean the two data sets and merge them to get the final consolidated output.

Below are list of operations and some snapshots performed to clean the two datasets.

**Dataset 1 : Energy dataset(rawdata.csv)**

The first step here was to read the rawdata.csv file in R. So we need to just work on the kWh unit, so we needed to remove the other units from the **rawdata**.



Once that was done we needed to **transpose the units and the time interval columns. We reshaped the data using the reshape library in R and applied the melting function.**

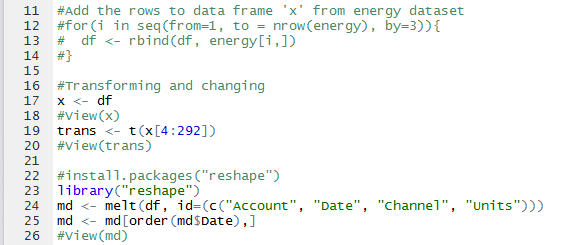


Fig: Code Snippet for melting & reshape the data

After this we restructured the data by **ordering it by date**. There was another issue with the rawdata that was provided. There was power outage provided for every 15 minutes throughout the day. So we aggregated the responses and narrowed it down to represent the power outages on an hourly basis for a particular day.

Next step was to acquire the **month, day, DayOfWeek and year** by breaking down the date. We achieved this by using the **lubridate** library in R.

On completion of these tasks we took a summary of the dataset which can be seen below:

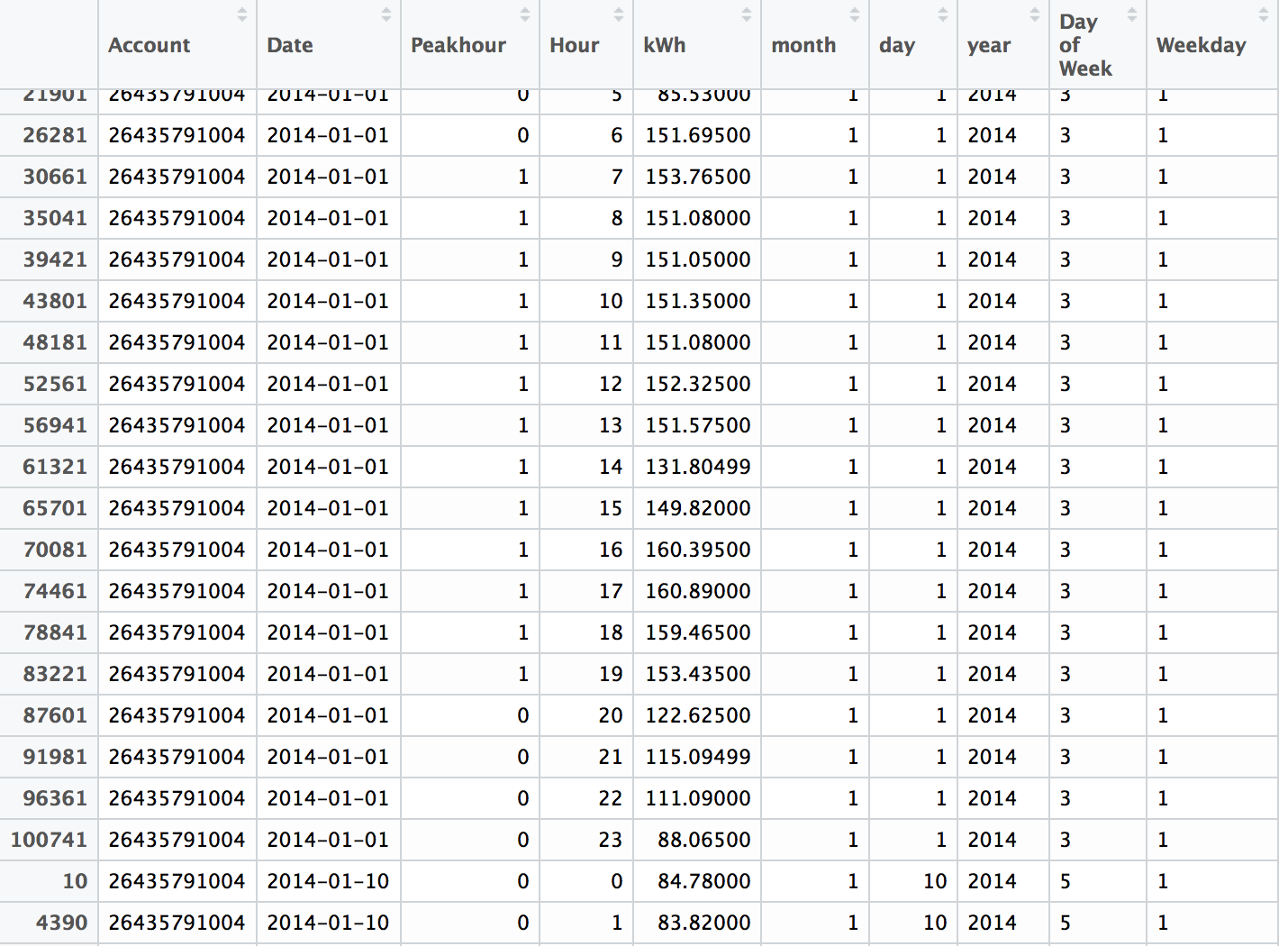


Fig : Snapshot of the restricted energy dataset

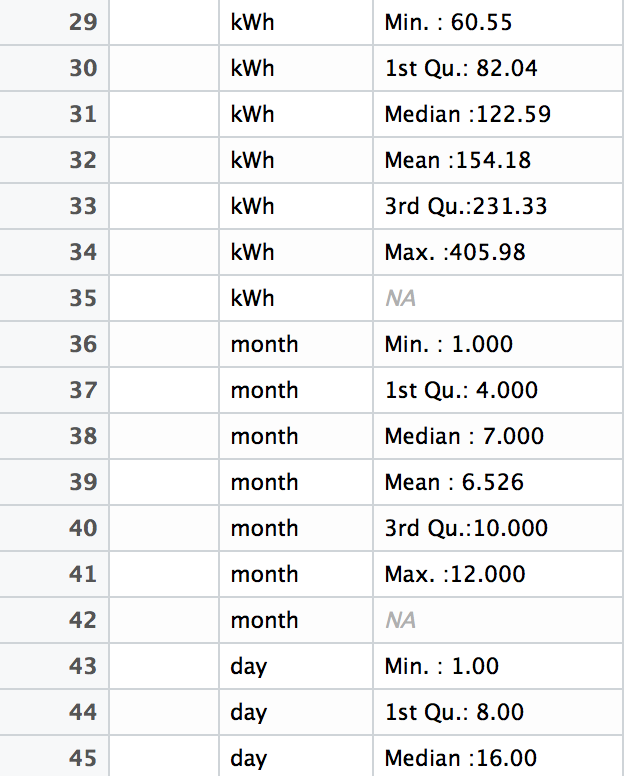


Fig: Summary for the energy dataset

Lastly we developed a function to detect the outliners and “NA” values in the given dataset and suggested a way to treat them Below is the code snippet for the function:

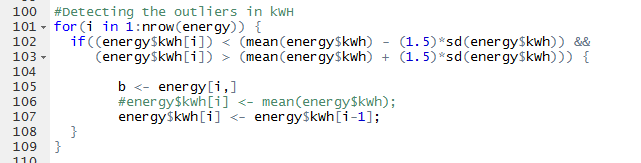


Fig: Code Snippet (Outliner function)

**Dataset 2 : Temperature Dataset**

On pulling the data from package of **WeatherData** in R, it is observed that there are multiple observations for a particular hour. We handled those observations by taking mean for that hour and then replacing the multiple data for that hour with the mean calculated.

For eg : For every temperature shift in an hour an entry is being made for the hour.

3.15 🡪 22F

3.35 🡪 21F  
3.55 🡪 23F

In this case the algorithm that we have written would take the mean of the temperatures and associate it to the corresponding R (3rd hour 🡪 22F)

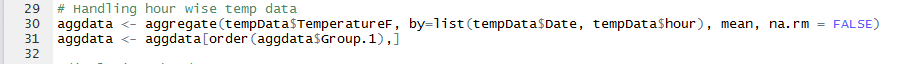
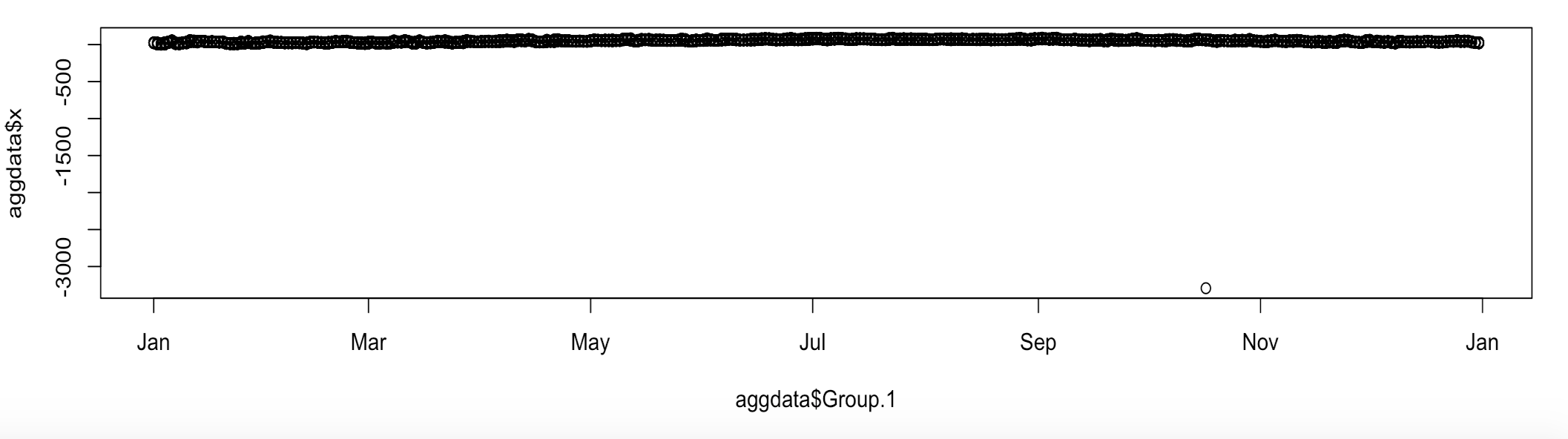


Fig : Code Snippet for handling multiple data for temp in an hour

We then proceeded with taking the summary and for the dataset and saw that there was a record for the temperature showing -324F which in this case was an obvious outlier. So we plotted the graph to check if there are more outliers.

**OutLiner Graph**

****

**Handling the outliners:**

In order to detect the outliners we wrote down a function and replace the outlier if any by the previous value in the column

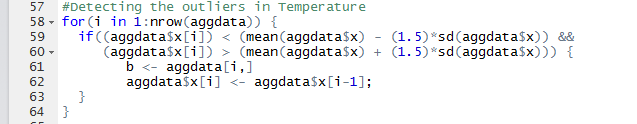


Fig : Code Snippet for Outlier function in Temperature dataset

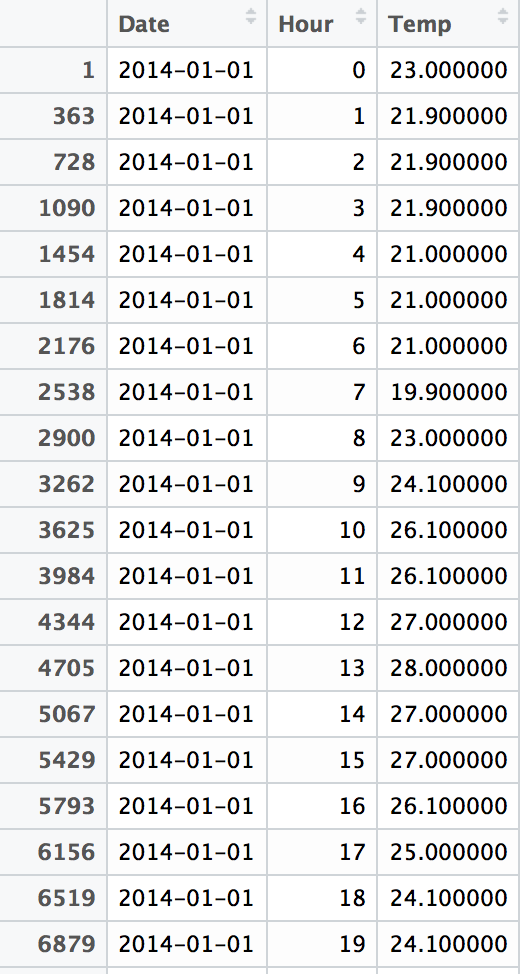
****

Fig: Final Temperature Dataset after cleaning

**Merging the two Datasets (Energy + Temperature 🡪 Consolidated )**

We finally merged the two dataset by using the dplyr package from R and applied the left outer join to achieve this. After this we took the summary for the merged dataset and it was found that there are multiple fields having “NA’s” and “0’s”

**Handling the NA’s in the merged dataset**

Now, we developed a function to detect the “NA”& “0” values in the given dataset and suggested a way to treat them Below is the code snippet for the function:

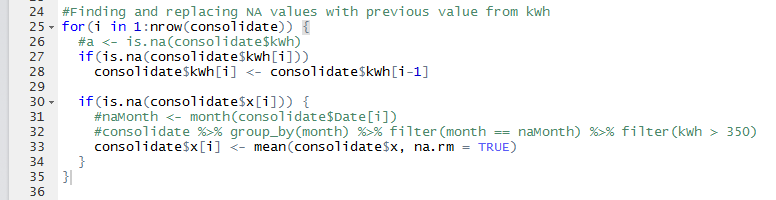


Fig : Code Snippet for handing NA’s & 0 values

**Testing for Outliners and replacing them:**

In order to detect the outliners we wrote down a function and replace the outlier if any by the previous value in the column

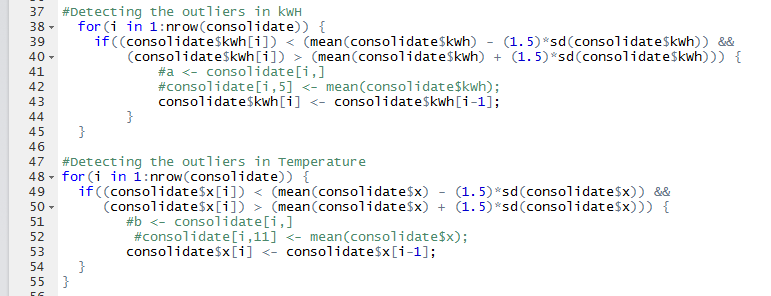
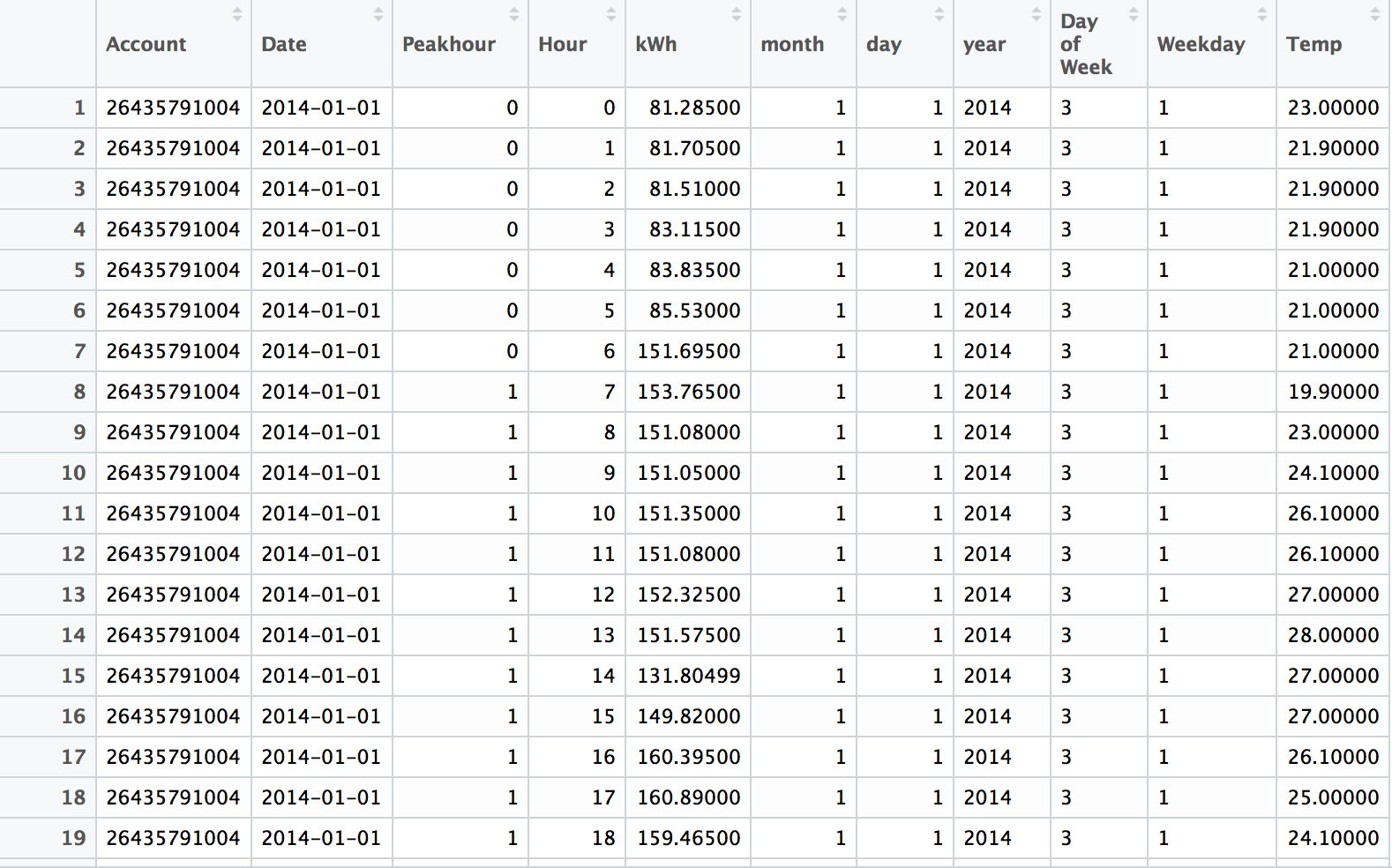


Fig : Code Snippet for handing outlier values

**Snapshot of Final Merged Dataset 🡪 Consolidated**



1. **Data Visualization Using Tableau**

With a view to get a wholesome idea about the dataset we decided to visualize it using a data visualization tool. The main objective of this exercise was to figure out which factors are dependent on each other and which ones are directly related to power consumption. Having some knowledge about these factors would help us to develop a better model which would thereby reflect on the prediction/forecasts made.

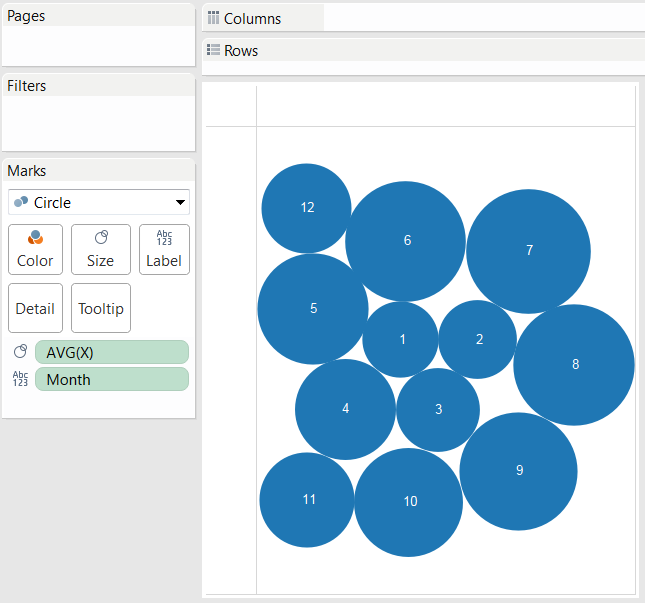
**HeatMap : PowerOutage\_byYear**



Fig : HeatMap : PowerOutage\_byYear

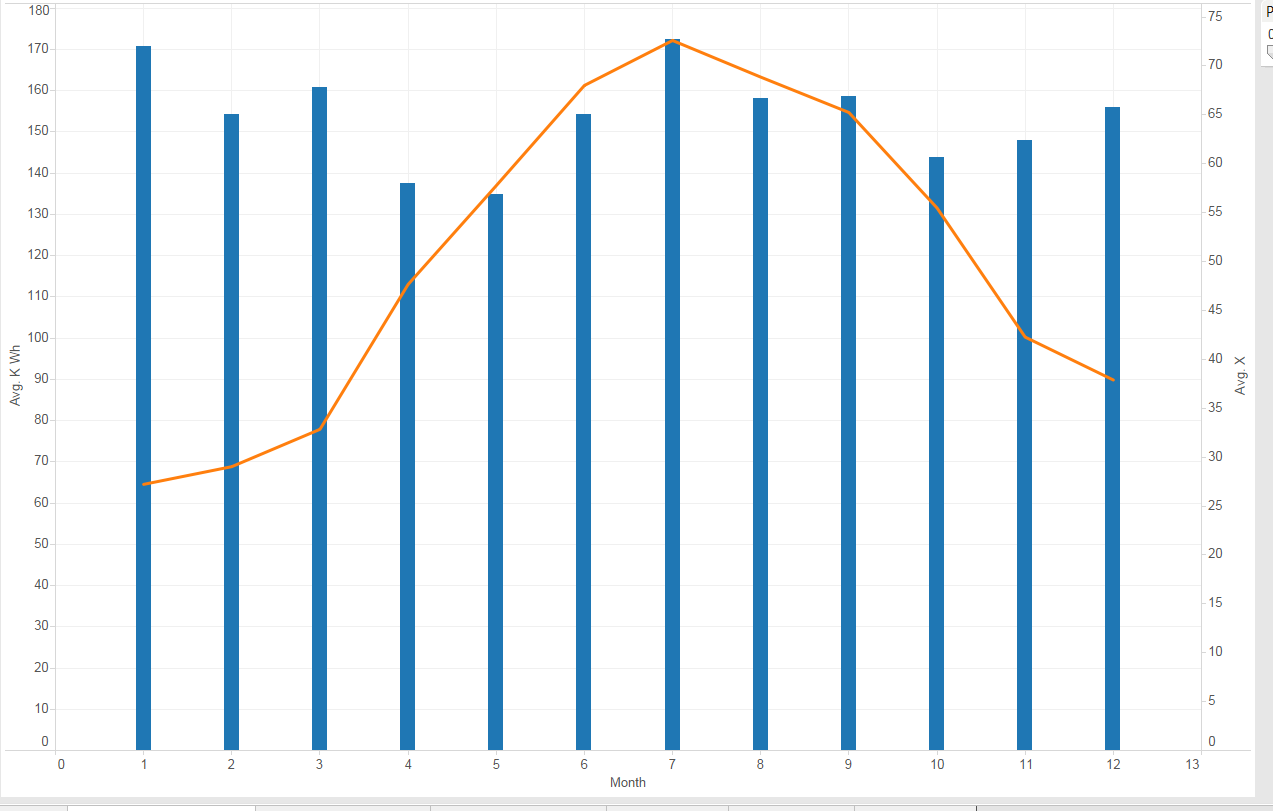
Observation: It can be seen that Power Consumption in the city was maximum in the month of July, January and March in the year 2014. This made us realize that temperature could be an effect for this increase.

**Temp Vs Month:**

****

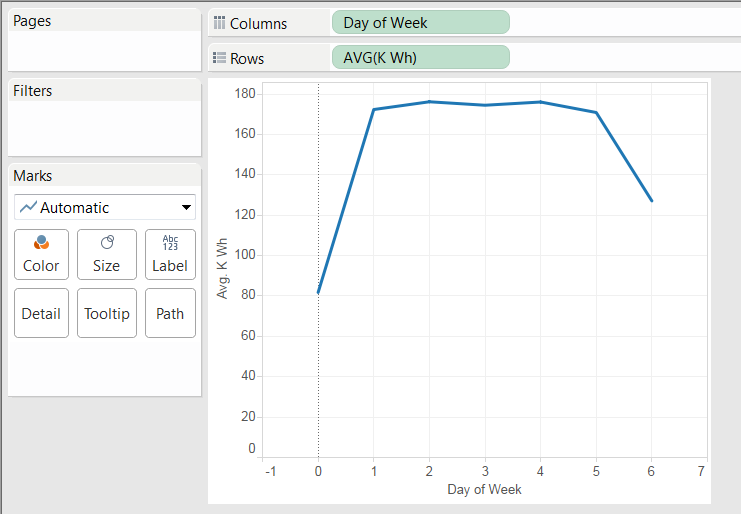
Observation: Here we can see that the temperature rises during the mid-year period around July, August and September and on the other hand it decreases during the end of the year around December and continues to the start around January,February and March. From this we can see that temperature has an effect on power outage. Cause the temperature in August is highest and so is the power consumption.

**Temp & Power Vs Month**

****

Observation: This visualization make our task a lot easier. Now we can see that temperature &month have an effect on the power outage. As the temperature decreases, the city of Boston uses more power (may be due to heaters and other devices). On the contrary as the temperature increases there is also a rise in power consumption (may be due to air conditioners, fans etc.)

**PowerOutage\_PerWeek**

****

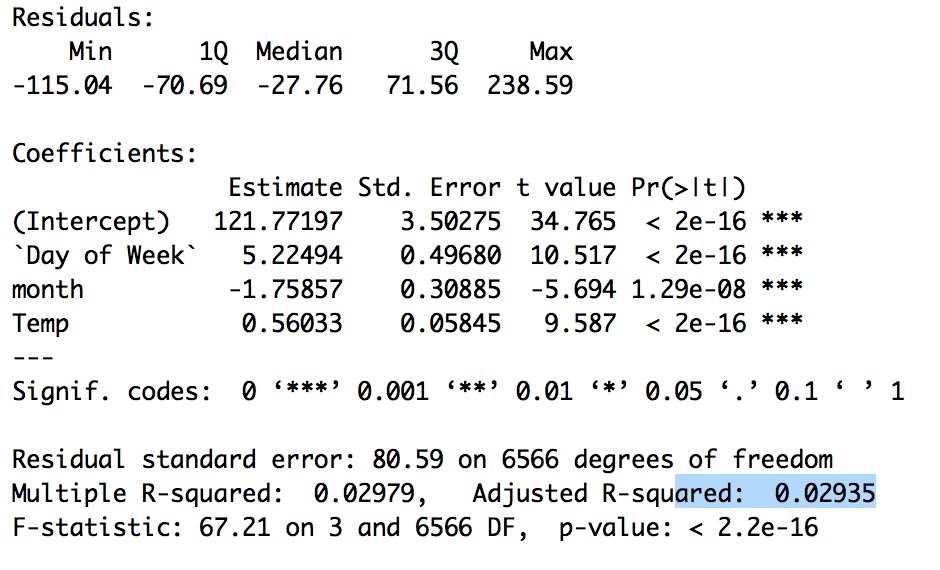
Observations: Here you can also see that day of the week also has an effect on power consumption. There is less power consumption during the start and end of the week. This can be explained by the fact that these two days are non-working days.

From all the observations we concluded that these are some of the factors that influence the power consumption are month, temperature, day of week.

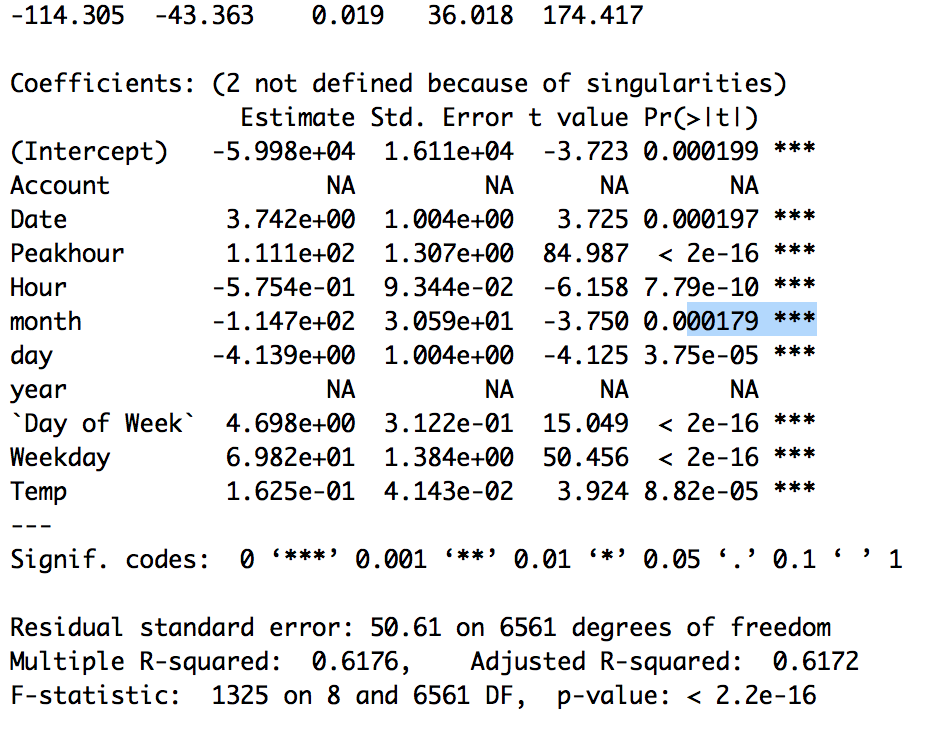
1. **Multi- Linear Regression**

In this stage we split the data up into training and test datasets in the ratio 0.75 : 0.25. Now we applied the linear regression model by selecting the response column as “kWh” and selecting the factors mentioned above i.e. month, temperature & day of the week.

**Performance metrics obtained on running the above model**

****

**Co-efficients, Std.errror, t-value & p-value**



As we can see from the above metrics there are a lot of other factors that also influence the power outage/consumption.

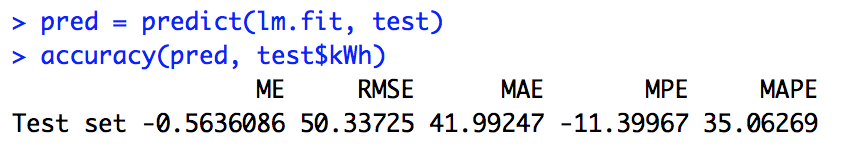
We have learnt in the lecture that p-values and t-value also play an important role in determining the factors. More the t-value more appropriate is the factor and similarly a factor having lesser p-value is more desired. As you can conclude that Peakhour and Weekday fit the above consideration so they need to be included in the model.

But on building the model it was clear that the adjusted r square value was not appropriate. So we decided to implement the forward selection regression method. In this method we basically tried to add one predictor at a time until the time that adding another predictor is no longer statistically significant.

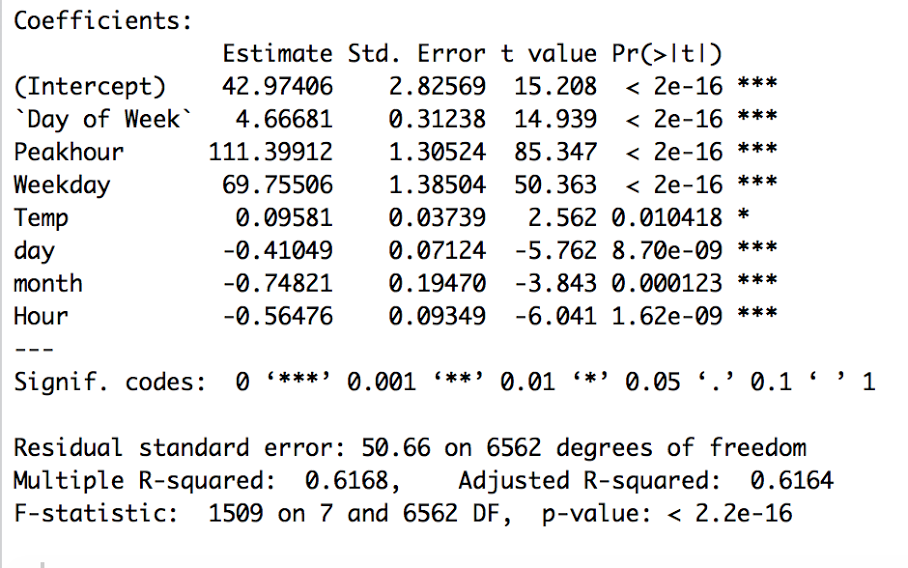
After many iterations of implementing forward selection we narrowed down the important factors affecting “kWh”. These are day, month, hour, Peakhour, Weekday and temperature.

Below is the snapshot for the evaluation metrics obtained on running the final iteration of the forward selection regression model.

**Performance metrics obtained on running the above model**



**Performance metrics obtained on running the above model**



1. **Forecast**

For the second part of the assignment the initial two parts are the same i.e. the data cleaning and multi-linear regression. On building the regression model, we observed that there are many 0 values for power. Because of this we are getting values for MAPE and MPE as infinity and –infinity. This is basically because we are getting a divide by zero scenario. In order to deal with this problem, we researched about this topic and then have applied log transformation function.

**What log transformation does?**

Takes the minimum positive number greater than zero . Let that be x. Replace 0’s in the dataset by log(x + x/2). In our case, that value is small as 0.015.

Based on this we have developed the model for forecasting the data. The performance metrics can be seen below:

1. **Refrences**

* Wunderground.com: <https://www.wunderground.com/weather/api>
* WeatherData : <https://cran.r-project.org/web/packages/weatherData>
* Power Factor: <https://en.wikipedia.org/wiki/Power_factor>
* A guide to reactive power-EDF Energy: <https://www.edfenergy.com/sites/default/files/b2b-guide_to_reactive_power.pdf>