**Part 1.1.1 and 1.1.2: Data wrangling and cleansing**

1. Dplyr and tidyr package is used.

Rawdata1.csv and Rawdata2.csv  is read in table format and data frame is created

Input: rawdata1 and rawdata2 dataframes

Filtering only the first kwh entries for each date

Kwh is aggregated hourly. Sum of 12 observations (5 min intervals rolled up to hourly)

Stack Data

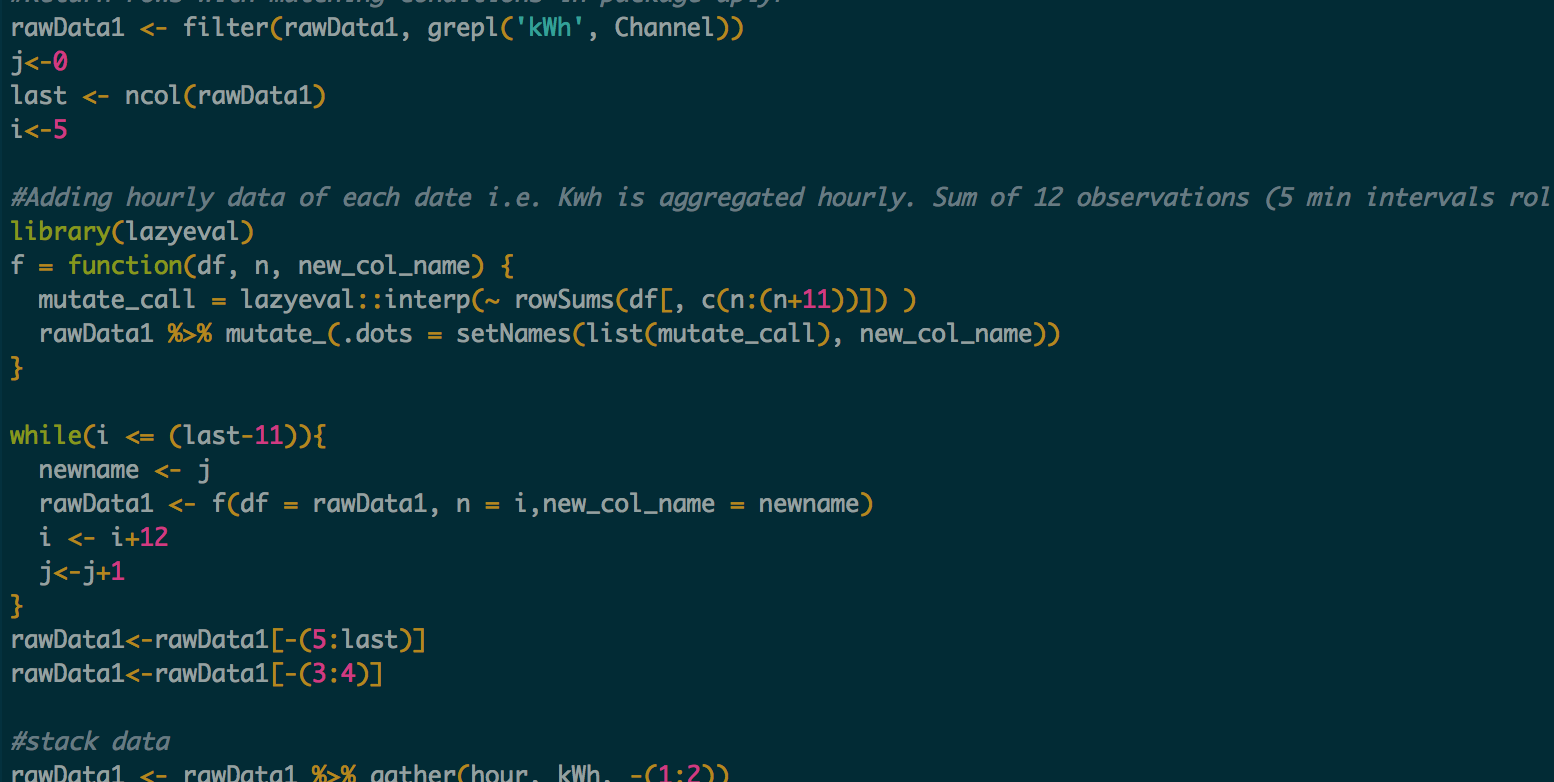
Weather data and raw data are joined by Date, hour

Output: rawdata1 and rawdata2 in sample format

Rawdata1 and Rawdata2 rows are binded

Output: Final data obtained

Unwanted values are marked as NA, Using locf package, missing values are filled, Rename all the wind direction with "Calm" and "Variable" as "north" as the wind direction degrees for north is 0







**Part 1.2: Multiple-Linear Regression**

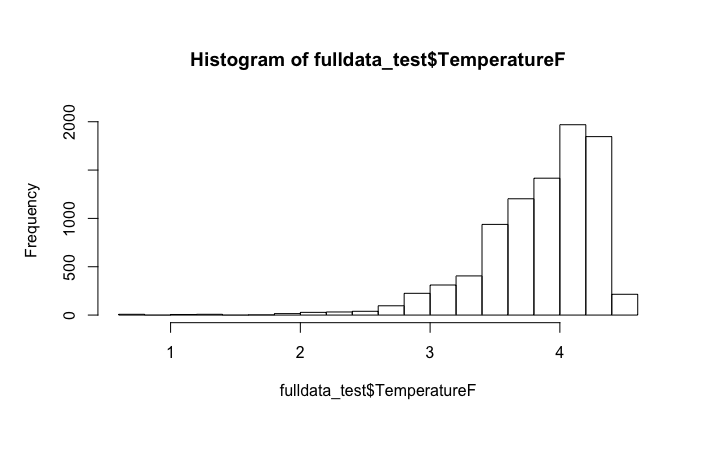
**Feature Transformation:**

log transformation on Temperature to make the growth approximately exponential.

This makes the data less skewed.

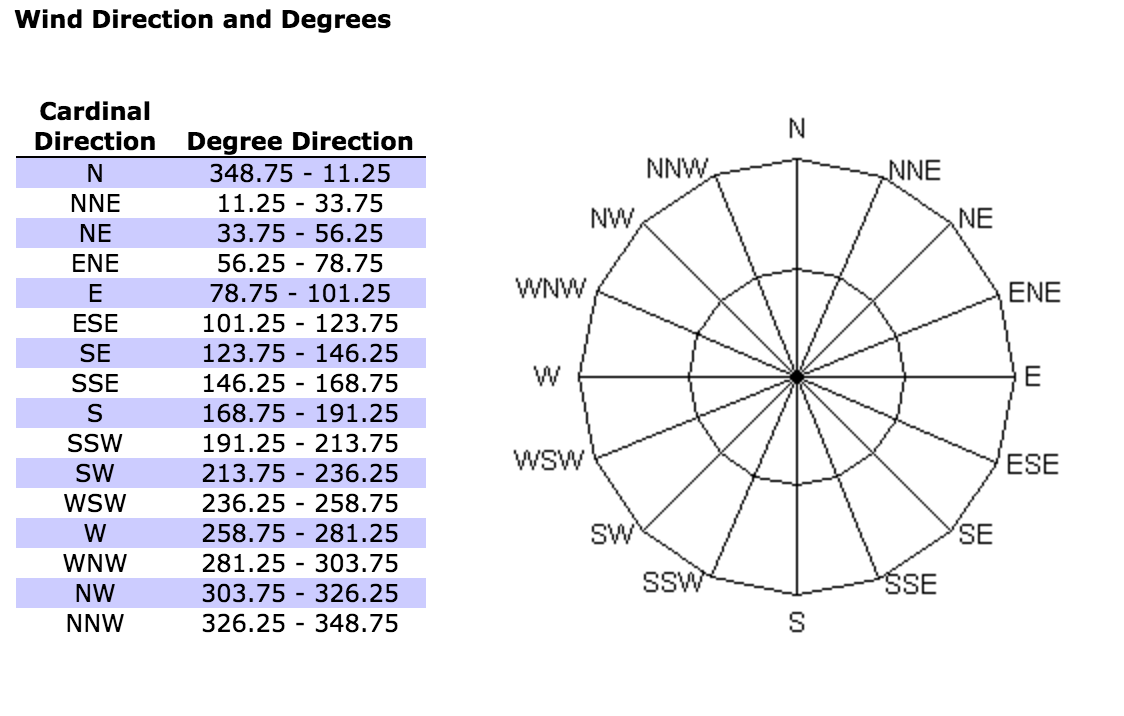
WindDirDegrees column: Introduced 2 variables, sin(WindDirDegrees) and cos(WindDirDegrees). This will yield less residual than using only WindDirDegrees

**Histogram of log(TemperatureF):**



**Feature Selection:**

* Date, Account, Year are not selected as they are single values
* Wind\_Direction is not selected as it is related to WindDirDegrees



* WindDirDegrees is not selected as it is split into cos(WindDirDegrees) and sin(WindDirDegrees). Hence, WindDirecSin and WindDireccos is selected
* Leaps package used for variable selection: Here subset regression is performed wherein number of subsets of each sizes is reported. Models are plotted and ordered by the selection statistic.
* Stepwise selection (forward, backward, both) is done using the stepAIC() function from the MASS package. stepAIC() performs stepwise model selection by exact AIC.

**Forward Regression:**

Data set , formula, nvmax = 14, method = “forward”

Regsubsets: returns separate best models of all sizes up to nvmax

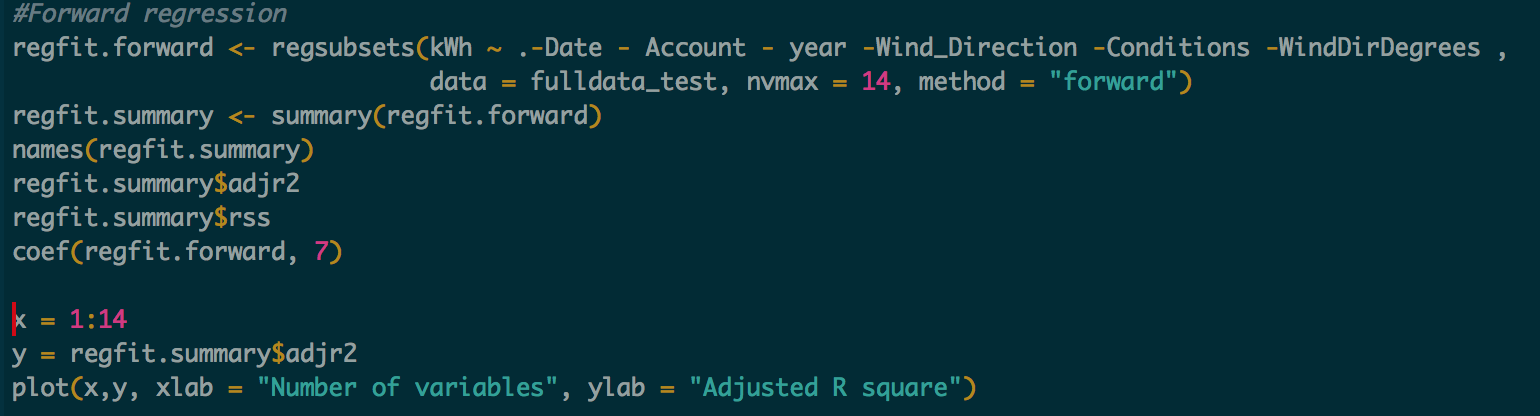
summary(regfit.forward): displays different models

regfit.summary$adjr2, regfit.summary$rss : Residual sum of squares for each model, Adjusted r-squared

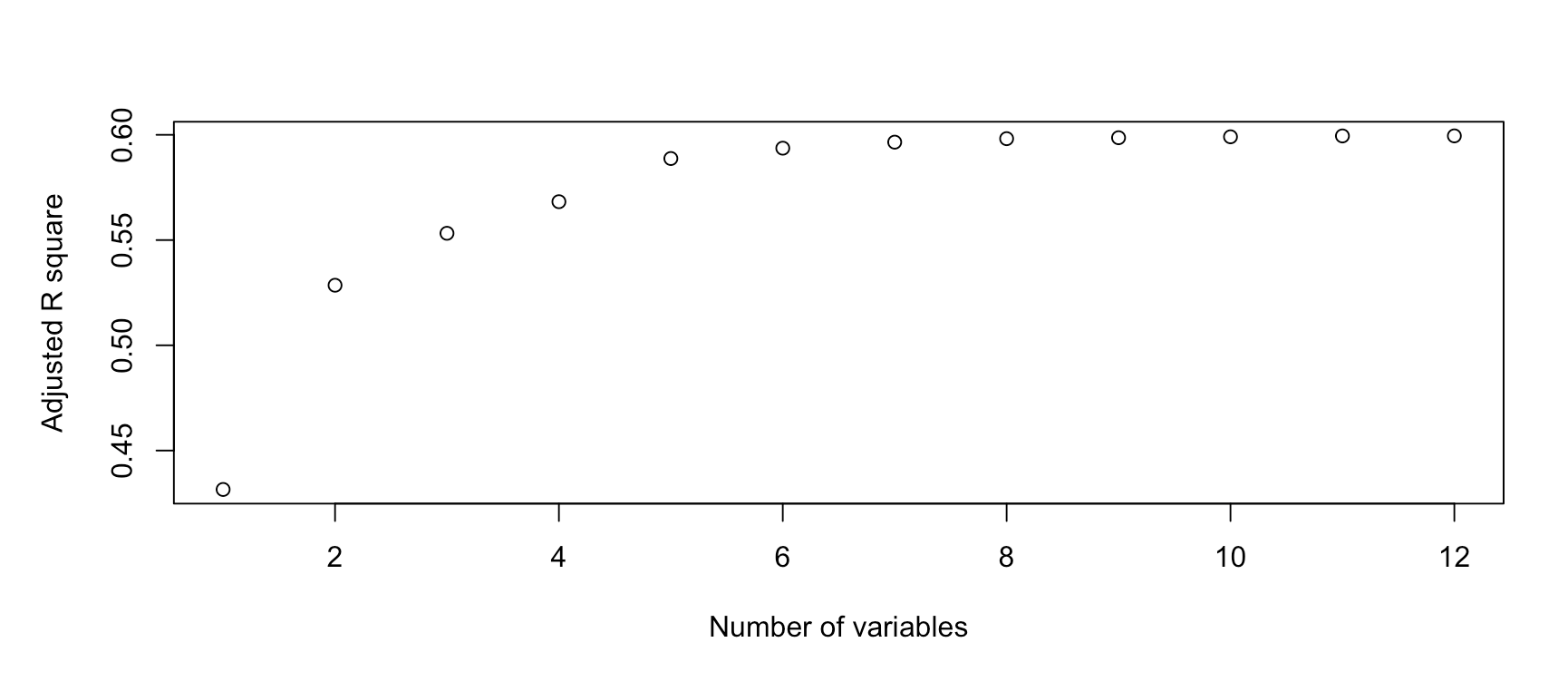
Output: Model coefficients are returned

Plot number of variables vs adjusted R square

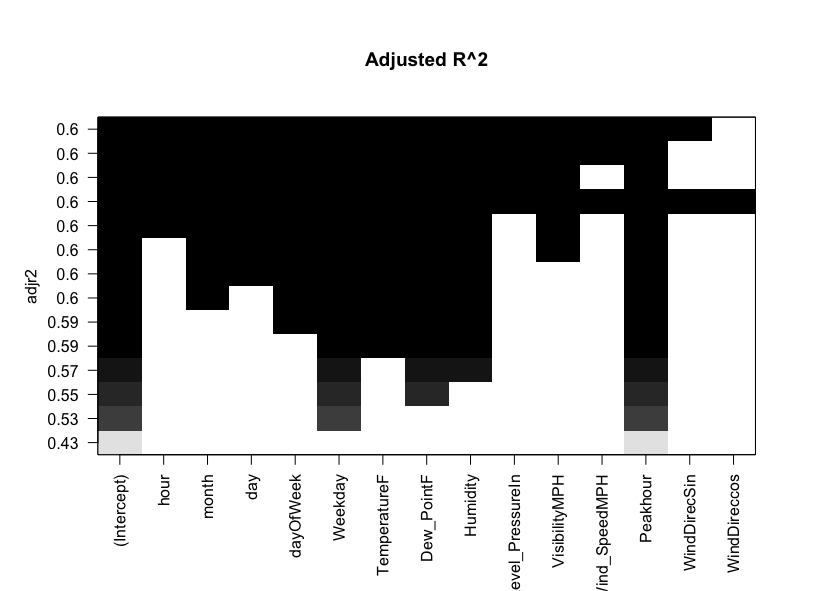
Adjusted Rsquare is constant after 7 variables. Thus, 7 features is selected using coef(regfit.forward, 7)







plot(regfit.forward, scale = "adjr2", main = "Adjusted R^2")



**Backward Regression:**

Data set , formula, nvmax = 14, method = “backward”

Regsubsets: returns separate best models of all sizes up to nvmax

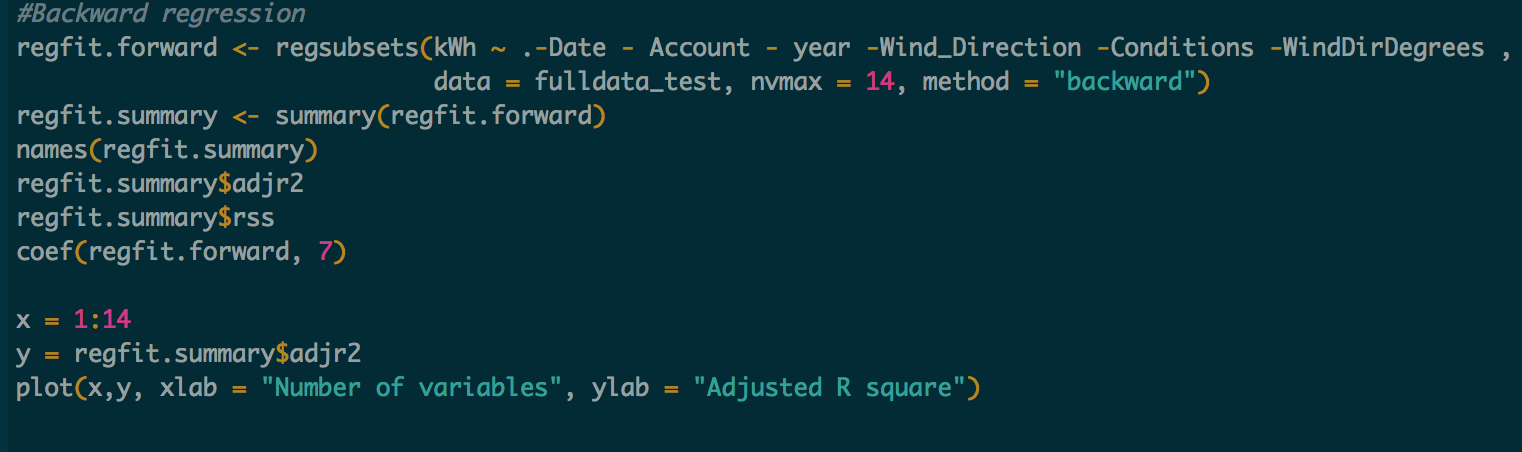
summary(regfit.backward): displays different models

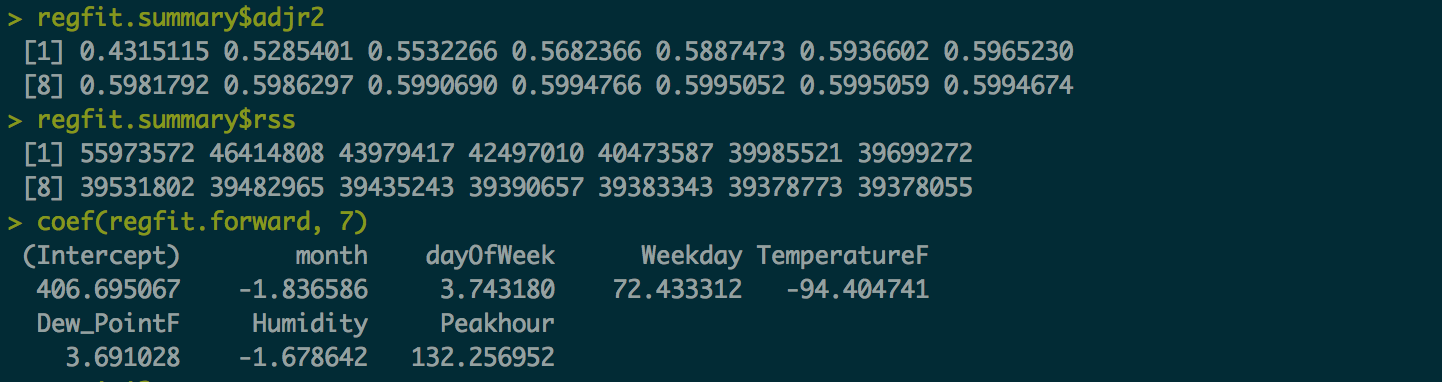
regfit.summary$adjr2, regfit.summary$rss : Residual sum of squares for each model, Adjusted r-squared

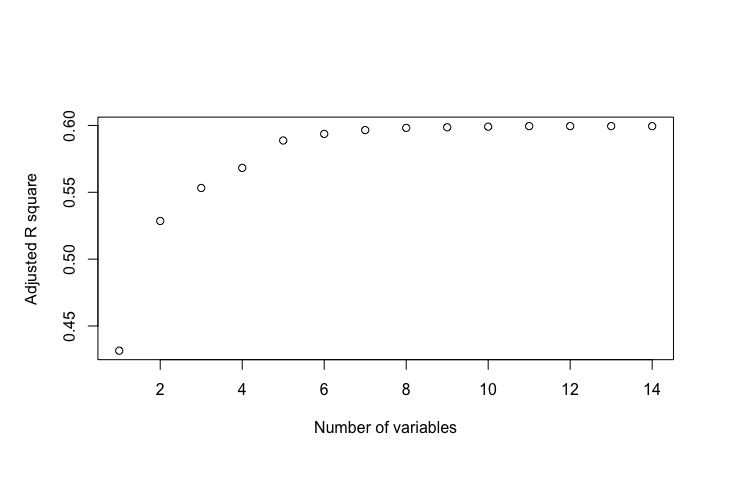
Output: Model coefficients are returned

Plot number of variables vs adjusted R square

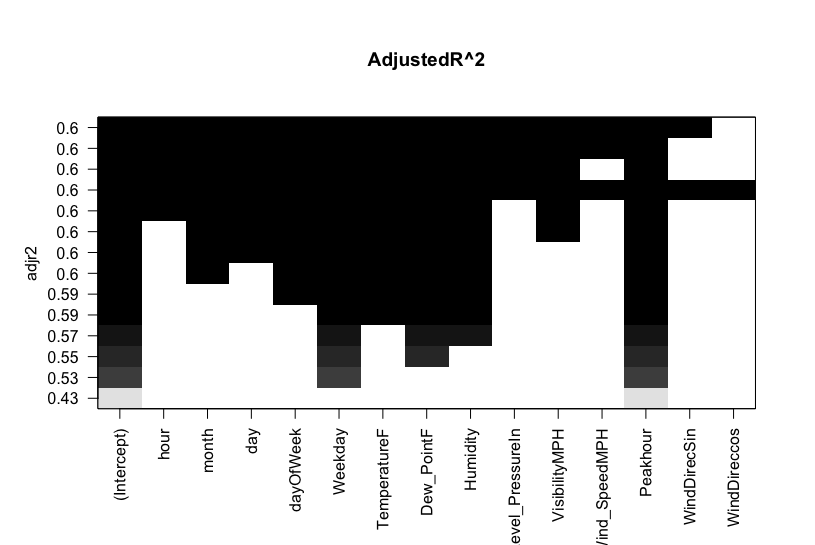
Adjusted Rsquare is constant after 8 variables. Thus, 8 features is selected using coef(regfit.backward, 8)







plot(regfit.backward, scale = "adjr2", main = "AdjustedR^2")



**Stepwise Regression:**

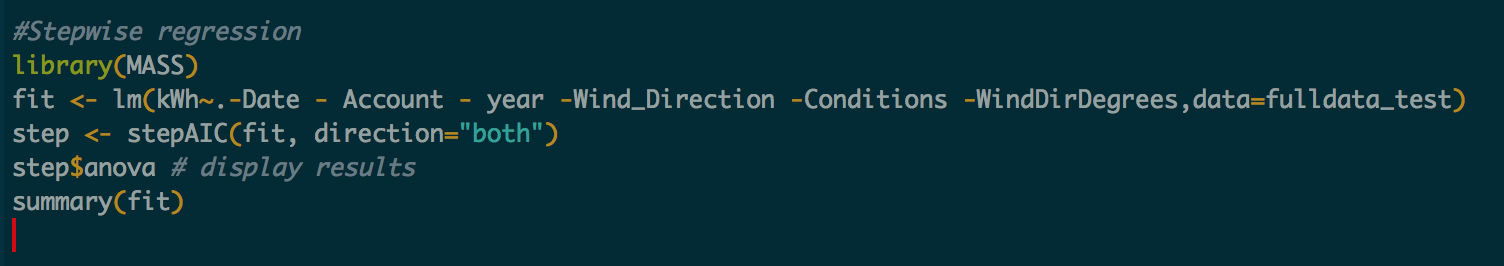
Data set , formula, using lm to fit linear models

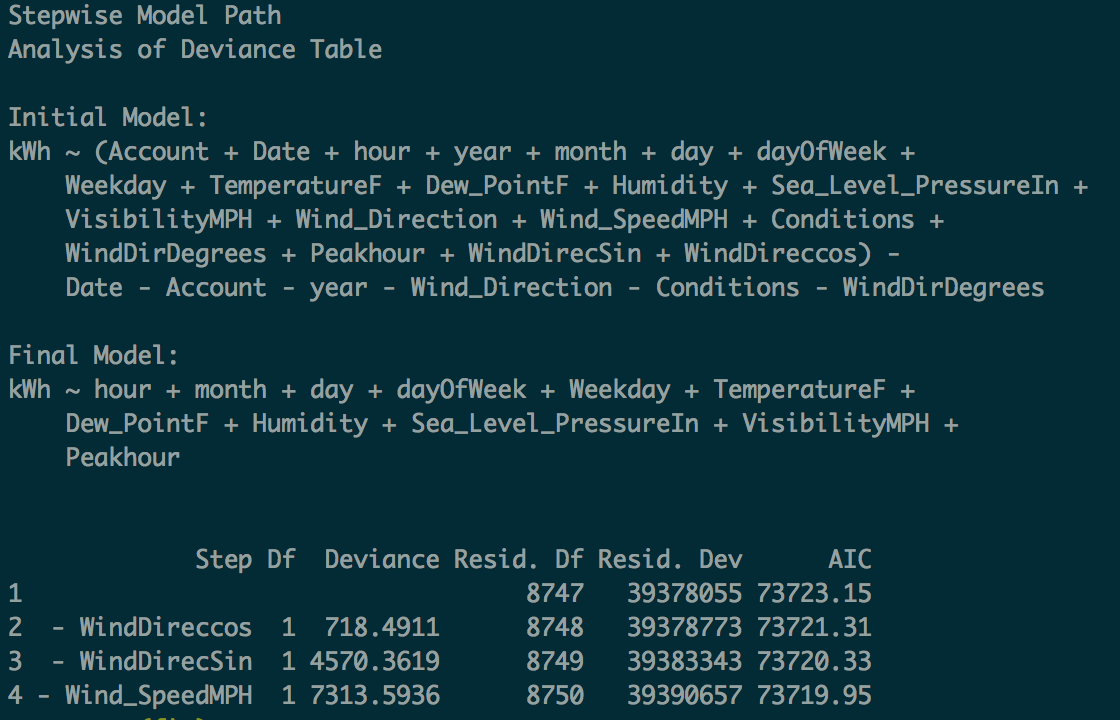
Choose a model by AIC in a Stepwise Algorithm:

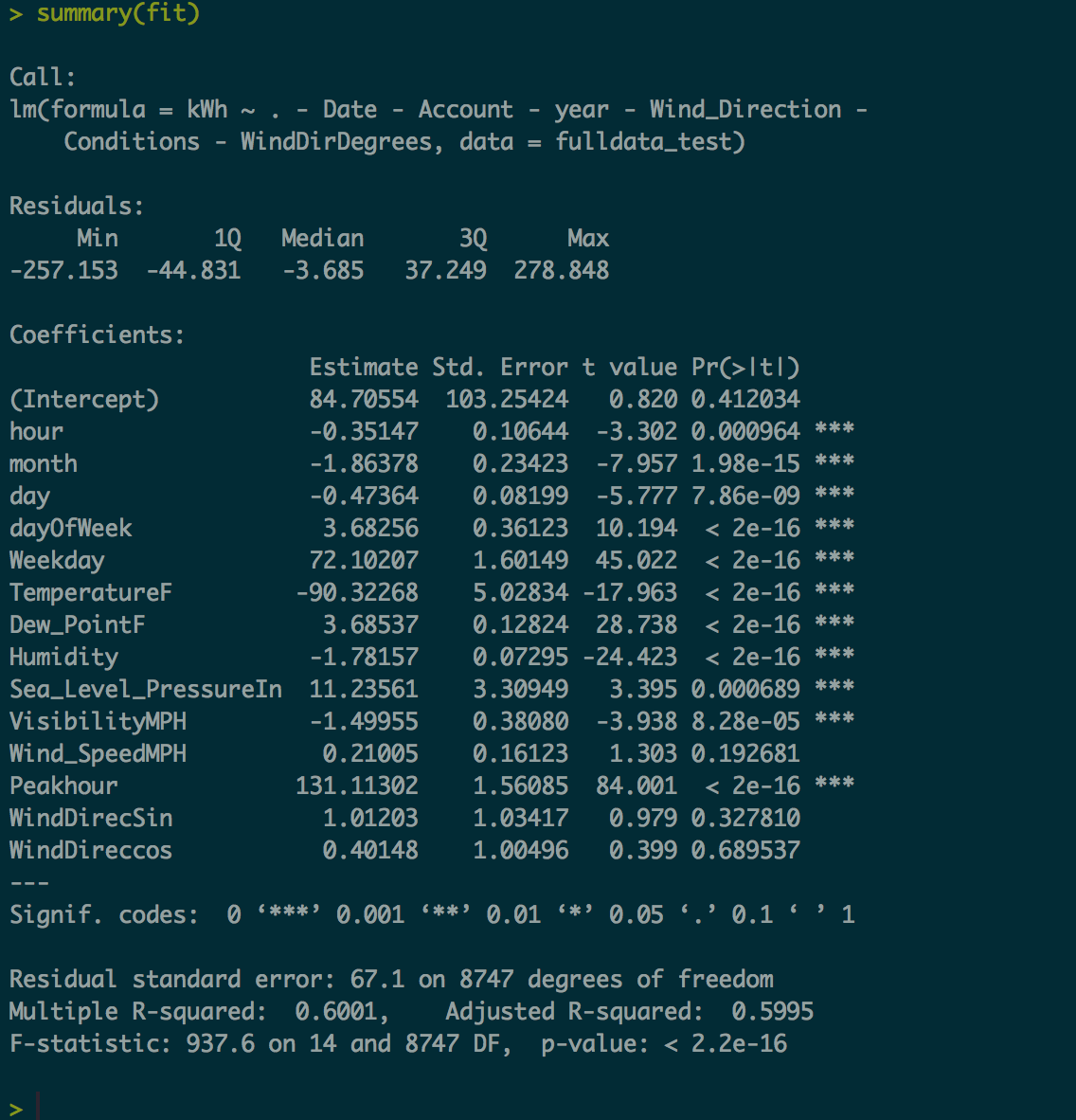
stepAIC(fit, direction="both")

Compute analysis of variance (or deviance) tables for one or more fitted model objects using anova.

Output: summary(fit) - WindDirecos, WindDirecSin and Wind\_SpeedMPH should not be selected as it has a high P value(>0.05)







**Model:**

Full data set, training data set, testing data set

Train 80% of model

Create a model with the trained set

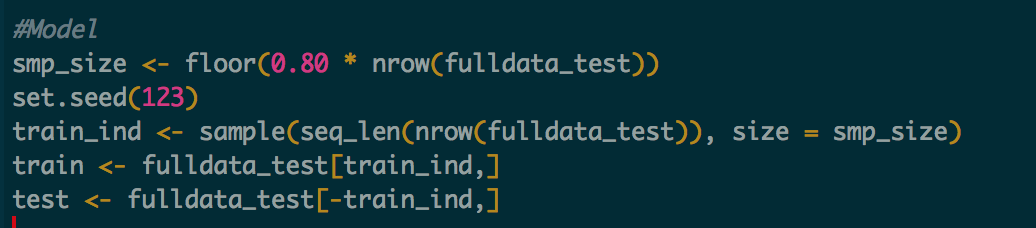
Output: summary of the model and plots

Model prediction of various model fitting functions

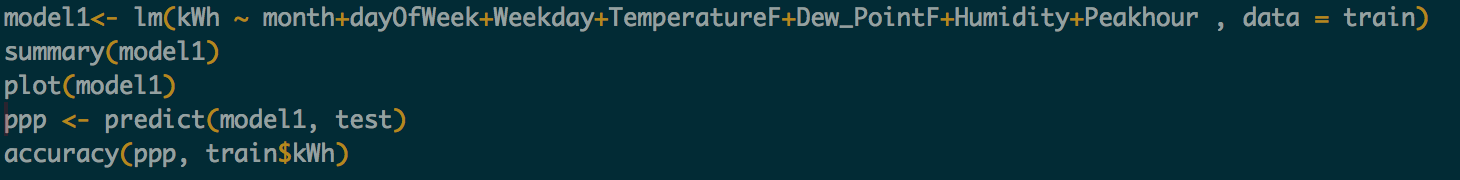
Output: RMS error, MAPE and MAE, Regression coefficients

Ridge Regression

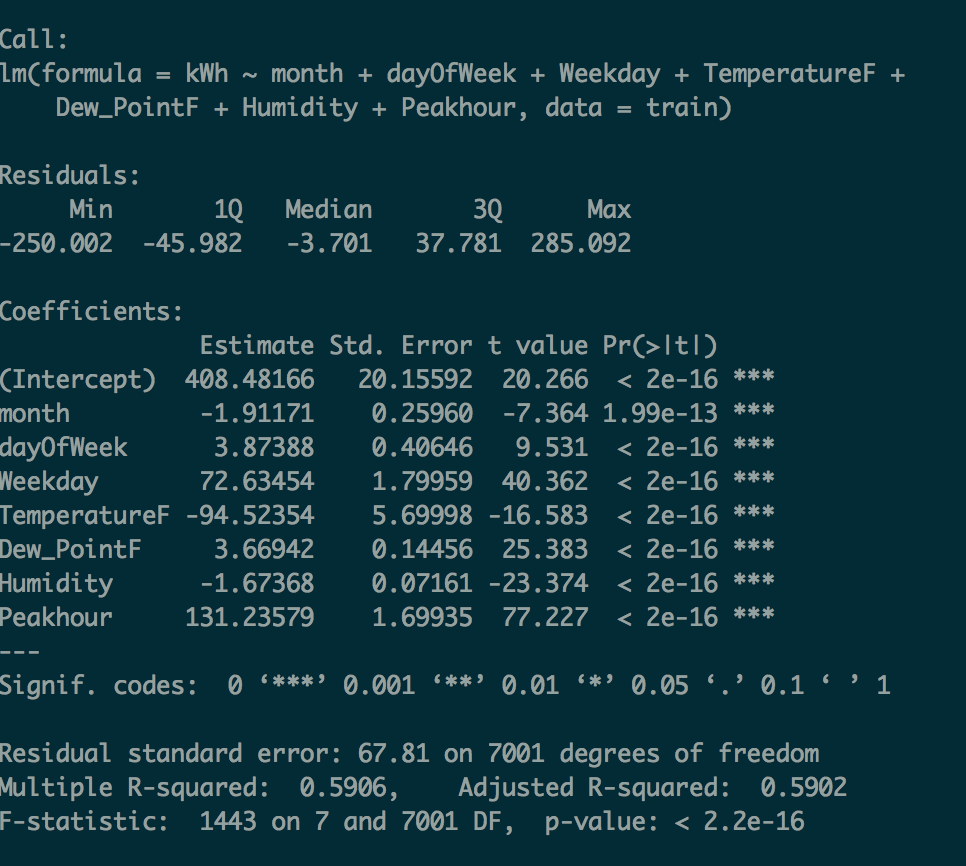
Training 80% model:



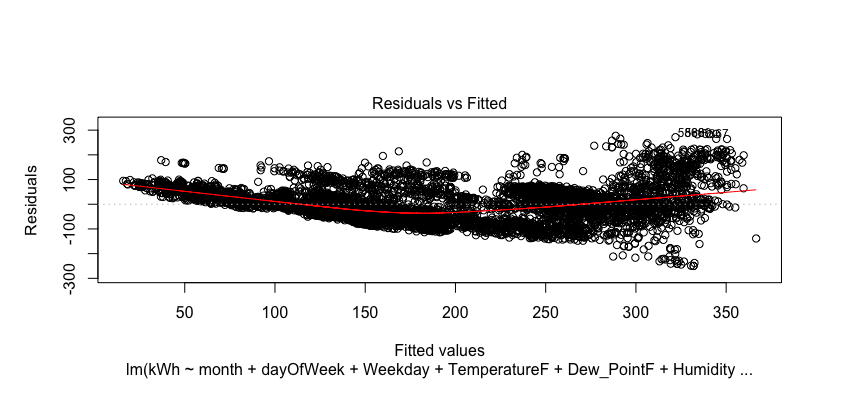
Creating a model:

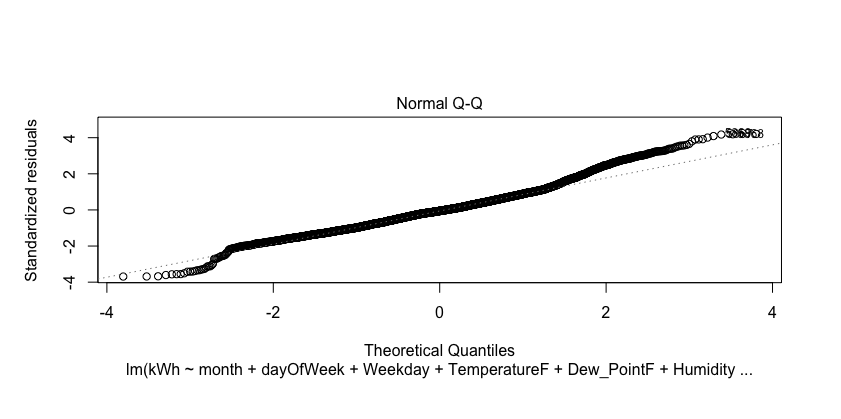


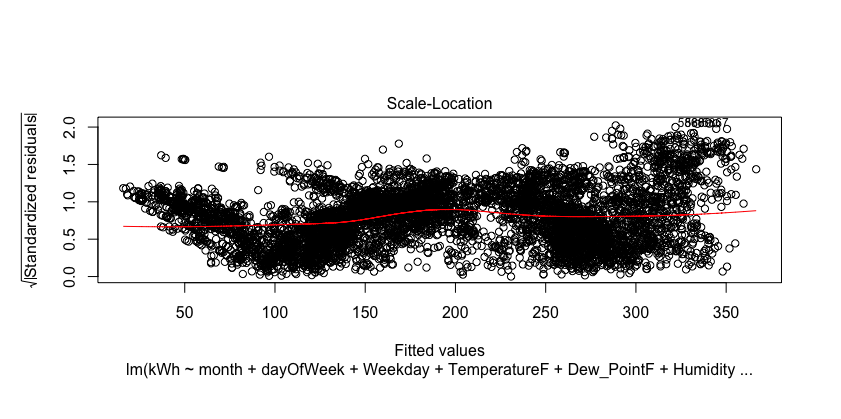
Below is the summary of model:

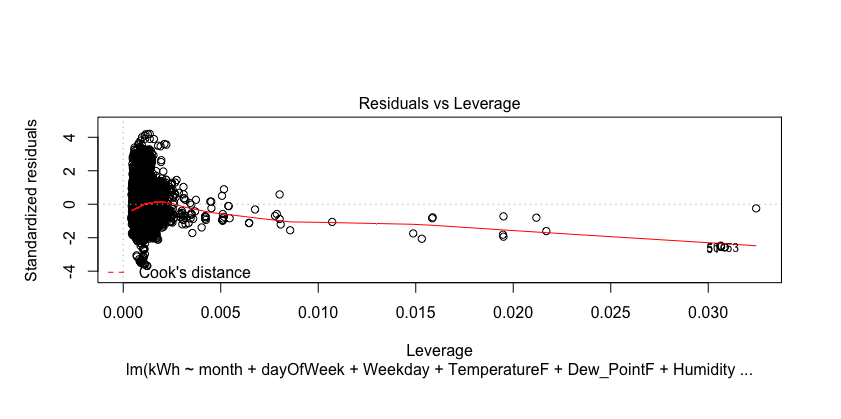


Plotting the model:

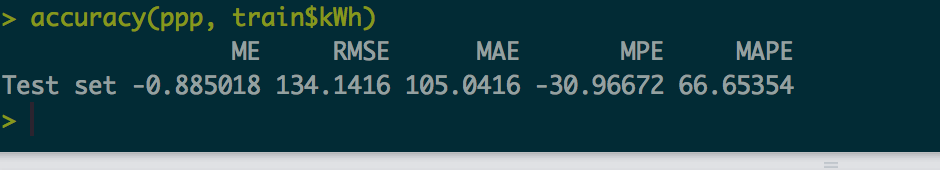






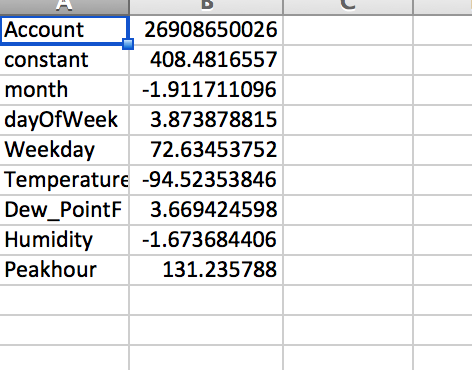


RMS error, MAPE and MAE:



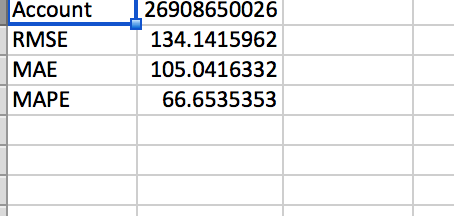
RegressionOutput.csv: using broom package



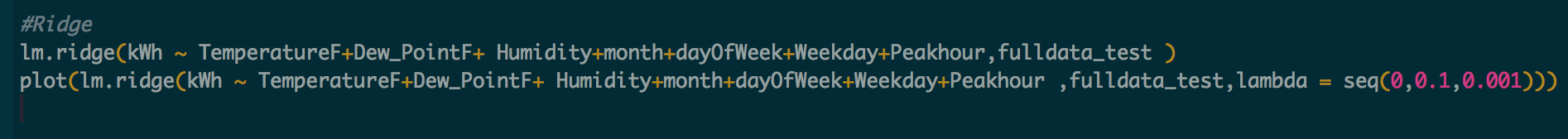


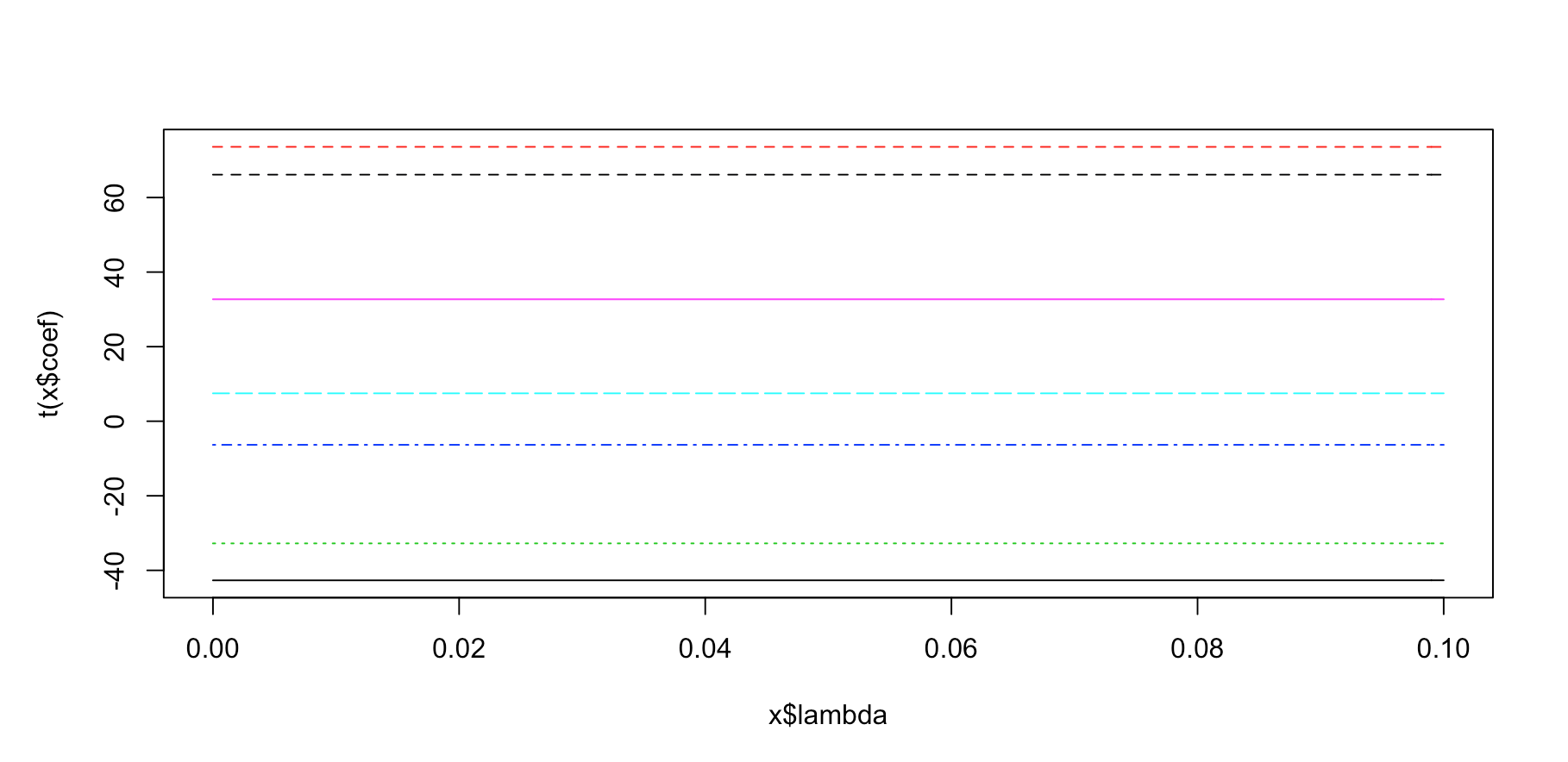
PerformanceMetrics.csv: using broom package





**Ridge Regression:**





There were large number of variables, high collinearity and hence adopted ridge regression.

Lasso will not necessarily yield good results in presence of high collinearity as the performance of the lasso will be dominated by ridge regression. Lasso selects only one variable among a group of predictors with high pairwise correlations.

**Part1.3 Forecast**

Input: Forecast data

Features: "TemperatureF", "Dew\_PointF", "Humidity" , "month", "DayOfWeek", "Weekday, "Peak\_hour" are selected

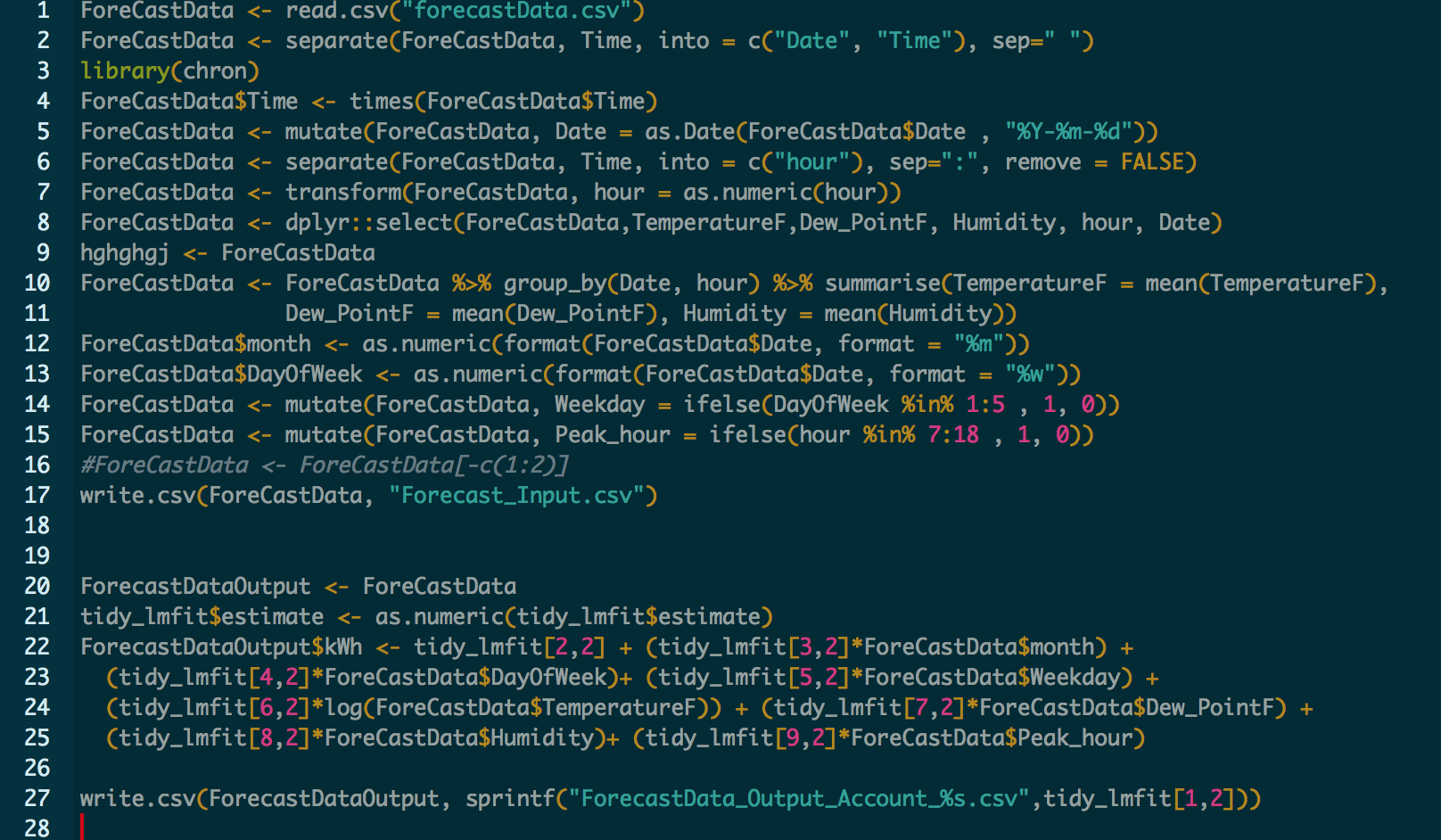
Output: ForecastInput data obtained

Data is forecasted using equation: kWh = Intercept + x1\*month + x2\*dayOfWeek + x3\*Weekday + x4\*log(TemperatureF) + x5\*Dew\_PointF + x6\*Humidity+ x7\*Peakhour

Where: Intercept = 408.4816557, x1 = -1.911711096, x2 = 3.873878815, x3 = 72.63453752, x4 = -94.52353846, x5 = 3.669424598,

x6 = -1.673684406, x7 = 131.235788

Output: ForecastOutput data



github link: https://github.com/pagarwal123/Assignment2\_Team12