CIND 820: Big Data Analytics Project

Section: DAH

Student Name: Muhammad Zaka Shaheryar

Student #: 500648718

Supervisor Name: Ceni Babaoglu

Date of Submission: June 19th, 2023

Title: Predicting Energy Demand in Spain

Table of Contents

1. **Abstract**

2. **Nomenclature**

3. **Literature Review**

4. **Data Description**

5. **Approach**

6. **References**

7. **Appendix**

1. ABSTRACT

Demand Forecasting is the important area for business and country alike. Energy demand is increasing due to increase in technological advancement. It is a need of time to tackle climate with machine learning (David Et Al.). Therefore, for this project the dataset that is chosen is

" Hourly energy demand generation and weather" (Kaggle). Dataset is taken from kaggle and it contains 35064 rows and 29 columns. The dataset have 4 years of electrical consumption, generation, pricing and weather data for spain. The data is retrieved from (Entsoe, Esios and Openweather). The links are given in references. The dataset have hourly data for electrical consumption and respective forecast by Transmission Service Opersator (TSO) such as Spanish esios Red Electric Espana (REE) for consumption and pricing.

The problem being considered for the project is to predict or forecast energy demand accurately in Spain. The research questions considered for this project are: 1. Which regression technique will accurately forecast the daily energy consumption demand using hourly period ? 2. How to accurately forecast energy demand 24 hour in advance compared to TSO? 3. Using classification, determine what weather measurement and cities influence most the electric demand, prices, and generation capacity? The tools that will be used are Python weka, Tableau, R, Excel and others as needed.

The systematic data analysis process approach will be used for the project. After data selection, initial analysis will be carried out followed by the exploratory analysis (EDA). Then experimental design and model building will be carried out. Finally performance evaluation will be done together with recommendations and conclusion.

2. NOMENCLATURE

AMI: Advanced Metering Infrastructure

ANN: Artificial Neural Network

ARIMA: Autoregressive integrated moving average

ARMAX: Autoregressive-moving-average model

BP: Back-Propagation

BPN: Back-Propagation Network

GBRT: Gradient Boosted Regression Trees

DNN:Deep Neural Network

GA: Genetic Algorithm

KPI: Key Performance Indicator

MAE: Mean Absolute Error

MAPE: Mean Absolute Percentage Error

ML: Machine Learning

MLP: Multi-layer Perceptron

MLR: Multiple Lineae Regression

MTLF: Medium-term Load Forecast

LSTM:Long-Short Term Memory networks

LTLF: Long-term Load Forecast

PCA: Principal Component Analysis

PDF: Probability Density Function

RMSE: Root Mean Square Error

RMSPE: Root Mean Square Percentage Error

RNN: Recurrent Neural Network

STLF: Short-term Load Forecast

SVM: Support Vector Machine

SVR: Support Vector Machine Regression

3. LITERATURE REVIEW

Tackle climate change with machine learning is motivation behind this project. The problem being considered for the project is to predict or forecast energy demand accurately in Spain. Accurate forecast of energy demand helps in advance planning and dispatching of resources which in turn save environmental cost due to extra production of Green house gases (GHGs) such as carbon dioxide to meet energy demand by buring fossil fuels. There are many opportunities to reduce GHG emission using ML as shown in figure 1 (David et al.).

There are many entities involved in energy market. Most important one are transmission system operator, power plant, commercial and residential users. **A transmission system operator (TSO)** (wikipedia) is an entity entrusted with transporting energy in the form of natural gas or electrical power on a national or regional level, using fixed infrastructure. The term is defined by the European Commission. The restructured electricity market operation is shown in figure 2 (Shahidehpour et al.). As can be seen, forecasting is necessary for ISO or TSO as well as GENCO or power plant. The research questions are reiterated in Table 1 and Githun link for the project is provided below https://github.com/Zaka123456/CIND-820

Table 1: Literature Review Research:

No.	Research Question (RQ)
1	**Which regression technique will accurately forecast the daily energy consumption demand using hourly period ?**
2	**How to accurately forecast energy demand 24 hour in advance compared to TSO?**
3	**Using classification, determine what weather measurement and cities influence most the electric demand, prices, and generation capacity?**

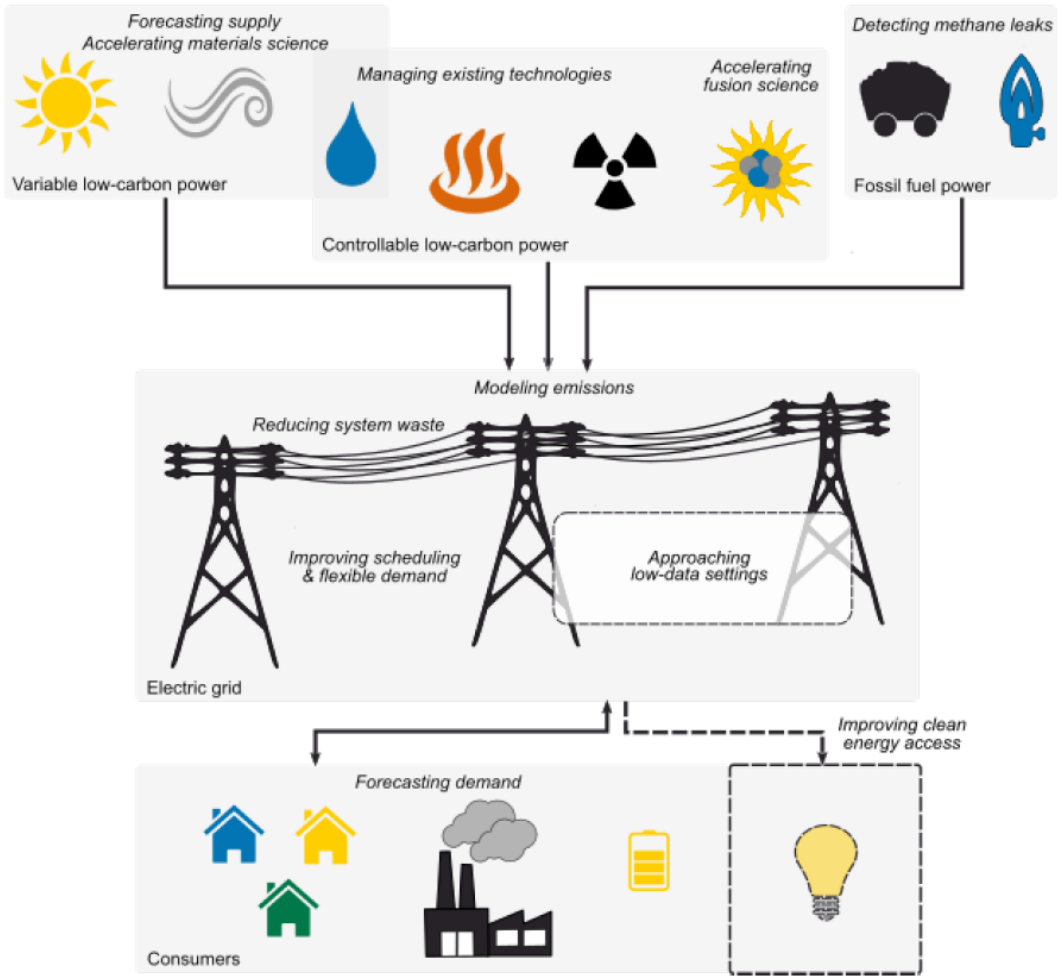


Figure 1: Selected opportunities to reduce GHG emissions from electricity systems using machine learning.

Figure 1: Selected opportunities to reduce GHG emission using ML

The general steps (Anton et al.) followed in review papers is shown in figure 3. As can be seen from the figure, ANN-based model is the most popular model in energy forecasting which is due to its non linearity in input to output matching. ANN is model inspired by human nervous system. ANN regression model comprises of input, weight, error, transfer function, activation function and output (Zakria et al.). The questions to considers for literature review are provided in Table 2. Six papesr are reviewed and key take-aways and how it can be useful to this project is provided below.

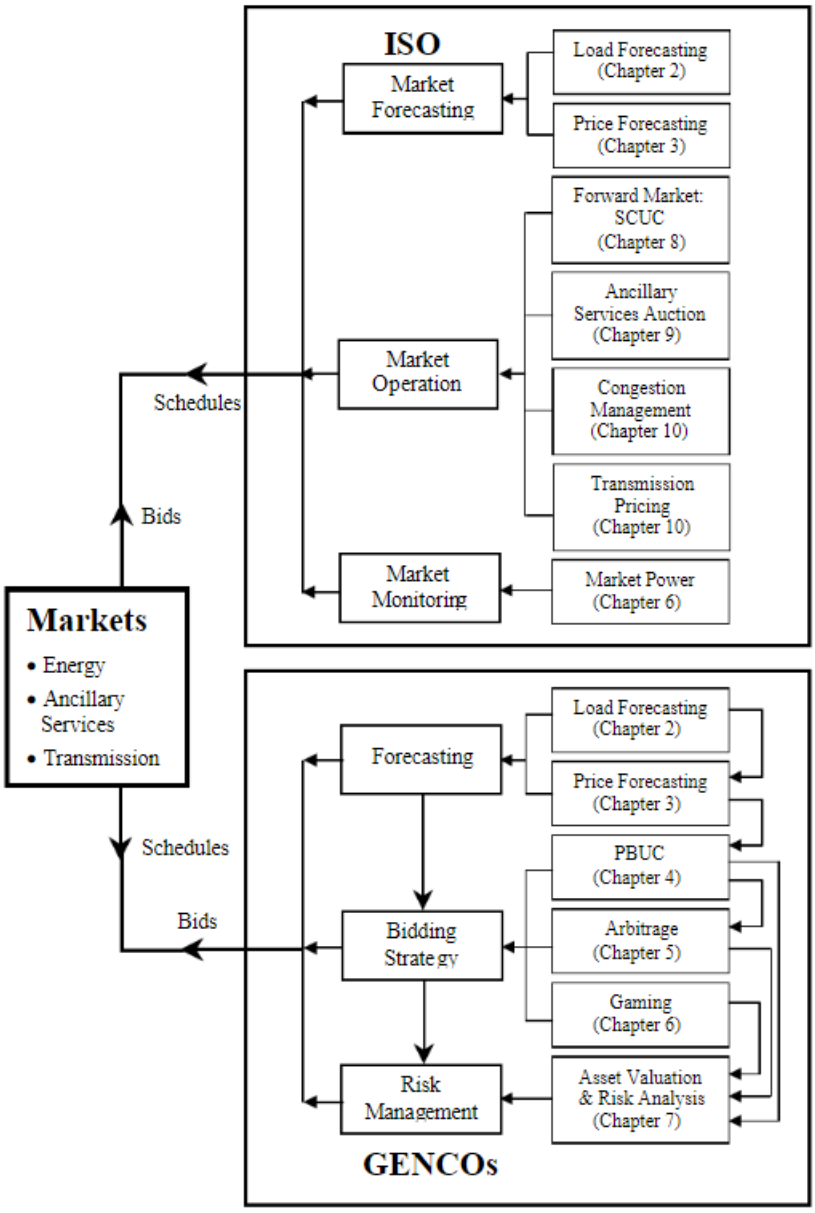


Figure 1.1 Restructured Electricity Market Operation

Figure 2: Restuctued electricity market operation

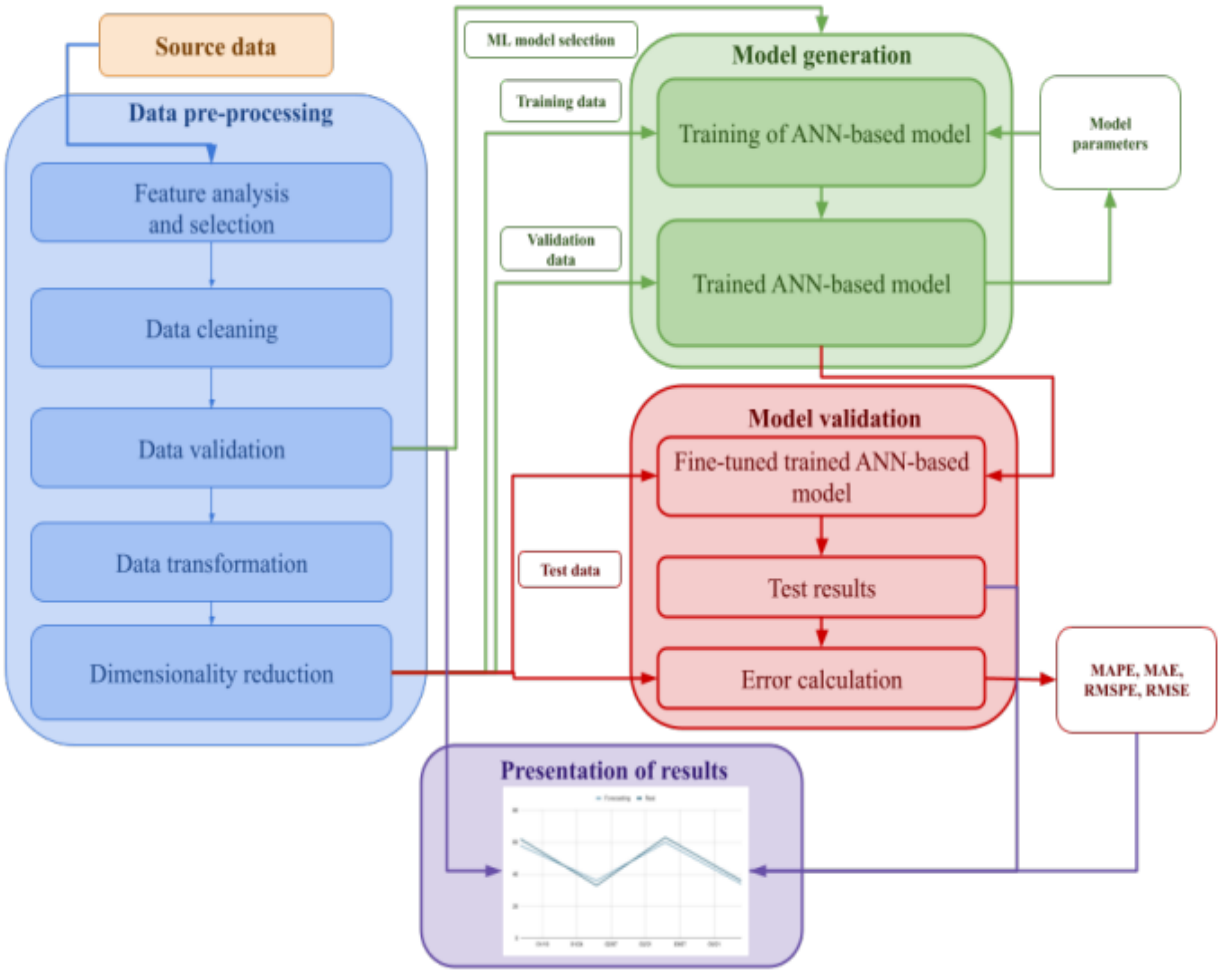


Figure 3: Generalization of steps documented in review papers

Table 2: Questions to consider for Literature Review

No.	Literature Review Question
1	What do you already know about the topic?
2	What do you have to say critically about what is already known?
3	Has anyone else ever done anything exactly the same?
4	Has anyone else done anything that is related?
5	Where does your work fit in with what has gone before?
6	Why is your research worth doing in the light of what has already been done?

Paper 1:

Yildiz et al. performed Regression and ML analyses on commercial building electricity load forecasting. The data is obtained from Kensington Campus and Tyree Energy Technologies Building (TETB) at University of New South Wales (UNSW). Importance of forecasting is mentioned in the paper together with brief review of Thermal models, and Auto Regressive models. Hourly electricity load data and minute interval weather data including Ambient Dry Bulb Temperature (DBT) and Relative Humidity (RH) is obtained from Campus and TETB electricity meters and local weather station respectively. Complimentary weather data is obtained from Sydney Observatory Hill Weather Station which is 7 km away from UNSW.

Comprehensive Regression analysis is provided. Input parameter are referred to as influence parameter as they influence the target parameter (load). Regression models and how to improve them is provided. Single vs multivariate regression models is considered. ML models considered are ANN, SVM, and Regression Trees. The authors analyze different ML algorithms for both campus and single building separately. Also they perform two type of Short term elecctricy load forecasts (STLF). First is day ahead Hourly forecast and second is Daily peak forecast. This is done for both buildings. Furthermore, data is analyzed for four seasons and for year 2013-2014. Summer months are December, January and February. Autumn months are March, April and May. Winter months are June, July and August. Spring months are September, October and November.

Influence parameters considered for hourly forecast are Previous day same hour load, Previous week same hour load, Previous 24 h average load, Working day/holiday binary indicator, DBT, RH, and Day of week. Similarly for peak load forecast; influence parameters are Previous day peak load, Previous week peak load, Previous week minimum load, Holiday/Business day binary indicator, DBT and hour of the day. For performance evaluation of models, the metrics used are R^2, R^2_{adj}, RMSE(%), MBE(%), and MAPE.

Results shows that ANN with Bayesian Regiulation Backpropagation is the best ML model. Furthermore, regression models performed fairly well compared to advanced ML models. Moreover, almost all model perform better prediction for overall campus load compared to single building load. Also, average day ahead hourly forecasts have higher accuracy compared to daily peak demand forecasts.

I can use the methods for improving regression models for my project that are discussed in paper such as Principal Component Analysis (PCA), Sensitivity analysis, and Stepwise regression.

Paper 2:

Young et al. proposed ANN model for forecasting sub-hourly (15 minutes interval) electricity usage in commercial buildings. They investigate even smaller unit than an hour for STLF. Data set for study is obtained from building management system (BMS) of a commercial building complex, and data are periodically pulled into a relational database. The site consist of three office buildings and they all are managed by onbe utility billing system. One main electric meter and several sub-meters are installed. The main meter measures electricity usage, both the instantaneous power in kW with minute interval and aggregated electricity usage at every 15 minutes in kWh.

Nine ML models considered are Simple Naive model, Gaussian process with radial basis function (RBF) kernel, Gaussian process with polynomial kernel, Linear Regression, ANN, SVM with normalized polynomial kernel, SVM with RBF kernel, K-Star classiifer, and Nearest neighbor ball tree. Significant predictor variables are previous electricity usage, Interval stamp (TIF), Day indicator (DTF), HVAC operation schedule (OPC), Outdoor dry-bulb temperature (ODT), and Outdoor relative humidity (ODH). For performance evaluation of models, the metrics

used are correlation coefficient, Coefficient of variance of the root-mean squared error or CV(RMSE), and Absolute Percentage Error (APE).

Results shows that ANN is the best algorithm for this study. For verification, two months (Augusta nd September) in year 2012 are used. Furthermore, three training methods are considered: Static, Accumulative and Sliding Windows. Cumulative and Sliding win dows train better than Static method. Therefore ANN with regularization algorithm for training is adopted. The model can provide a day-ahead electricity usage profile with sub-hourly intervals and daily peak electricity consumption with a reasonable accuracy.

Paper 3:

Mucahit et al. proposed a time series forecasting- based peak shaving algorithm for building energy management. Peak shaving helps to reduce the peak electricity demand resulting in reduced cost for end-users. A smart building is any structure that uses a central controller to automatically regulate energy demand. The peak load of the system is the highest amount of energy consumption during the day that is characterized by short time periods. To handle peak demand, peak load shaving is an attractive strategy. It involves using battery energy storage systems (BESS) where the secondary energy storage device allow a microgrid utility to shave the peak demand by chagrin g the BESS when demand is low, and discharging it when demand is high.

The publicly available dataset provided by the U.S Department of Energy on household electricity consumption is used for analysis. The data is taken from March 6, 2011, to April 5, 2012. The peak of electricity on weekdays is higher than on the weekend. Furthermore, peak of electricity occurred between 11 pm and 1 am on weekdays and between 12 pm to 3 pm on weekends. Moreover, for classifying predicted load, electricity loads are classified into seven

groups. First class (C^1) is associated with range from 0 to 1000 kW electricity, and last class (C^7) is associated with range from 5500 kW or higher electric consumption. C^1 is best region for recharge action while C^7 is best region for discharge action.

The ML models used in regression analysis for one-step ahead forecasting are as follows: Naive Baseline (taking load/label previous week load as predicted value for today), Random Forest (RF), Light Gradient Boosted Machine (LGBM), and Long-Short Term Memory Netwrorks (LSTM). Performance metrics for regression analysis are Normalized deviation (ND), Normalized Root Mean Squared Error (NRMSE), and MAPE. Furthermore, ML models used in classification analysis for load forecasting are as follows: K-Nearest-Neigbours with Dynamic-Time-Wrapping (KNN-DTW), RF, XGBoost, LGBM, LSTM, Fully Convolutional Networks (FCN), and Residual Networks (ResNet). Performance metrics for classification analysis are Accuracy, F1-score, Precision, and Recall.

The results shows thats RF is the best algorithm for both regression and classification tasks. The RF is then used to develop a forecasting-based peak shaving algorithm, which first predict the peak period, and these predictions are then used to determine the decision of charging and discharging the battery. I can apply the knowledge of ML models used in study since I have both regression nd classification analysis in my project.

Paper 4:

Kody et al. performed Heating, Cooling, and Electrical Load forecasting for a large-scale district energy system. The study covers a large-scale district energy system that simulateouly produce electricity, heating and cooling for a large University of Texas campus at Austin. The Hal C.Weaver power plant and associated facilities provides all the cooling, heating and electrical needs for the campus. It is connected to city grid but only used in case of emergency. The load

profiles for cooling and electric loads fluctuate significantly while that of heating load remains almost constant for summer-time condition for example in August. Furthermore, Cooling load fluctuate more compared to electrical load. In winter-time condition, e.g in February, heating load profile increased while cooling load profile decreased compared to summer-time condition. Moreover, in winter, heating load fluctuates much more making it difficult to forecast them accurately.

The input parameters are dry bulb temperature and RH (weather variables). The correlation analysis in the study provide useful insights regarding relations between input varaibles and three different loads and also among three different loads as well. I can also apply the visualization similar to figure 4 in study as it give clear pictures which variables are related to other strongly. Each load 9Electrical, Cooling and Heating) is strongly correlated to ambient dry bulb temperature and less so with ambient relative humidity. As expected, cooling and electrical loads are positively correlated to temperature and negatively correlated to humidity, while the reverse is true for heating. Furthermore, all the loads are highly correlated with each other suggesting they all undergo similar variations.

ANN is still the dominant methodology for forecasting building energy loads. Moreover, time series analysis models are an improvement upon ANNs in case of time dependency of data points e.g, energy load forecasting. Time series analysis models are categorized into two groups, time-domain and frequency-domain. Therefore, three models namely ANN, Linear Auto Regressive with eXogenous input (ARX), and Nonlinear ARX. Furthermore, all three models are developed first using weather input only and then both weather and time inputs. Coffectient of determine (R^2) is a performance metric. The results shows that for one day ahead prediction for cooling, heating and electrical loads, NARX with both weather and time variables outperform all other models.

Paper 5:

Zakria et al. performed energy output forecasting analysis of Hybrid photovoltaic (PV)-wind system using feature selection technique for smart grid. The motivation is to enhance smart grid by efficiently predicting energy produced by renewables energy system. Weather factors have significant impact on power output of hybrid system. The weather factor (input parameters) selected for study are Solar irradiation (Solar power per unit area), Wind Speed, Ambient Temperature, Humidity, Precipitation, Atmospheric pressure, and wind direction. Solar irradiation, Wind Speed, Ambient Temperature, and Humidity have most significant impact on PV-wind power output. Target feature can be either power or energy. Goal is to determine energy demand trend over long period.

Historical hourly weather dataset is gathered using calibrated sensors at Middle East Technical University, North Cyprus Campus from Jan 1st to Dec 26th, 2015. Seven regression models are used in study namely: Extra tress Regressor, AdaBoost, SVR, K-Neigbors Regressor, Gaussian Process Regressor, MLP Regressor and Linear regressor. Performance metrics are MSE, MAE, R^2, and computational time. All seven models are first compared without any feature selection technique and then using Recursive Feature Elimination using Cross Validation (RFECV) technique. Feature Selection method such as RFECV can improve overall computational efficieny.

Results shows that Linear Regression is the best model for both analysis with and without feature selection technique. Also analysis showed that attributes are linearly dependent on each other. Development model stages (Figure 4) and Pair Plot (Figure 7) are together useful information that I can apply in my project. Other information I learn is that for every 1 degree in temperature, there is 5% decrease in RH. Furthermore, For RH > 50%, retaltionship between temperature and RH become linear.

Paper 6:

Navin et al. performed analysis to improve the forecasting model for predicting solar generation based on multiple weather metrics or input parameters. The analysis is similar to paper five above as it involve renewable energy source for power generation. Electricity renewables generates are not easily predictable and they varies based on both weather and site specific conditions. Therefore, focus of study is to automatically generate models that accurately predict renewable generation using historical weather forecast. Ten months monitoring period is used (Jan-Oct) in 2010. Data is divided into 80%-20% train-test split. Historical forecast data is obtained from National Weather Service (NWS) available at www.weather.gov and observational solar intensity data (in (W/m^2) is obtained from University of massachusetts Amherst weather station. The study is the extension of previous study by same author where they just focus on one input metric (sky cover) to predict solar intensity. Since intensity depends on many factors, therefore this study explore other factors apart from sky cover as well. These factors are temperature, dewpoint, windspeed, Precipitation potential, RH, and specific day of year. To ease the analysis, study focus on predictions at noon.

Four Models are analyzed and compared for training as well testing data using RMSE as performance metric. The models considered are: Simple Past Predict Future (PPF), Linear regression with sky cover as independent variable, Linear regression with multiple independent variables, SVM with Radial Basis Function (RBF) kernel and four principal components. STLF i.e, 3 hour in future is provided by the models.

The results shows that SVM-RBF with four principal components is the most accurate model compared to other models both in training and testing data. The second best is Linear regression with multiple independent variables. I can use the knowledge of Principal component analysis (PCA) provided in the study to improve the ML model in my project.

4. DATA DESCRIPTION

Dataset Description

- " Hourly energy demand generation and weather" (Kaggle) is the chosen dataset.
- Dataset is taken from kaggle and it contains 35064 rows and 29 columns.
- The dataset have 4 years of electrical consumption, generation, pricing and weather data for Spain.
- The data is retrieved from (Entsoe, Esios and Openweather). The links are given in references.
- The dataset have hourly data for electrical consumption and respective forecast by Transmission Service Opersator (TSO) such as Spanish esios Red Electric Espana (REE) for consumption and pricing.
- The attributes type is given in table 3. Here chr mean character and num means numeric.
    - Out of 29 attributes, 26 are numeric, 3 are non-numeric
    - Out of 3 non-numeric attributes, 2 are logical and 1 is character
- Out of 26 numeric attributes, 6 attributes have either zero or missing values. These attributes are provided in table 4. They can be removed from analysis as there is no information regarding these attributes.
- Uing R and python, descriptive Statistics for remaining 20 numeric attributes is provided in Table 5
- For regression analysis, **total.load.forecast** (row 17 in table 5) is selected as dependent variable.
- For classification analysis, **forecast.wind.offshore.eday.ahead** is selected as dependent variable

Table 3: Attributes Type

Attribute	Type	Attribute	Type
time	chr	generation.nuclear	num
generation.biomass	num	generation.other	num
generation.fossil.brown.coal.lignite	num	generation.other.renewable	num
generation.fossil.coal.derived.gas	num	generation.solar	num
generation.fossil.gas	num	generation.waste	num
generation.fossil.hard.coal	num	generation.wind.offshore	num
generation.fossil.oil	num	generation.wind.onshore	num
generation.fossil.oil.shale	num	forecast.solar.day.ahead	num
generation.fossil.peat	num	forecast.wind.offshore.eday.ahead	logical
generation.geothermal	num	forecast.wind.onshore.day.ahead	num
generation.hydro.pumped.storage.aggregated	logical	total.load.forecast	num
generation.hydro.pumped.storage.consumption	num	total.load.actual	num
generation.hydro.run.of.river.and.poundage	num	price.day.ahead	num
generation.hydro.water.reservoir	num	price.actual	num
price.actual	num		

Table 4: Numeric attributes with zero or missing values

Numerical Attribute
generation.fossil.coal.derived.gas
generation.fossil.oil.shale
generation.fossil.peat
generation.geothermal
generation.marine
generation.wind.offshore

Dataset Description continued

- Using R, Box plots for 20 non-zero numeric attributes are provided in Appendix
- Most of box plots shows that attributes are normally distributed which need to be further explored.
- Correlation Matrix is provided in figure 4

Table 5: Descriptive Statistic of Numerical attribute with non-zero values

No.	Arrtibute	Mean	St dev	Min	Max	NA
1	generation.biomass	383.5	85.4	0	592	19
2	generation.fossil.brown.coal.lignite	448.1	354.6	0	999	18
3	generation.fossil.gas	5623	2201.8	0	20034	18
4	generation.fossil.hard.coal	4256	1961.6	0	8359	18
5	generation.fossil.oil	298.3	52.5	0	449	19
6	generation.hydro.pumped.storage. consumption	475.6	792.4	0	4523	19
7	generation.hydro.run.of.river.and. poundage	972.1	400.8	0	2000	19
8	generation.hydro.water.reservoir	2605	1835.2	0	9728	18
9	generation.nuclear	6264	839.7	0	7117	17
10	generation.other	60.23	20.2	0	106	18
11	generation.other.renewable	85.64	14.1	0	119	18
12	generation.solar	1433	1680.1	0	5792	18
13	generation.waste	269.5	50.2	0	357	19
14	generation.wind.onshore	5464	3213.7	0	17436	18
15	forecast.solar.day.ahead	1439	1677.703	0	5836	0
16	forecast.wind.onshore.day.ahead	5471	3176.313	237	17430	0

Table 5 continued

17	**total.load.forecast**	**28712**	**4594.101**	**18105**	**41390**	**0**
18	**total.load.actual**	**28697**	**NA**	**18041**	**41015**	**36**
19	**price.day.ahead**	**49.87**	**14.619**	**2.06**	**101.99**	**0**
20	**price.actual**	**57.88**	**14.204**	**9.33**	**116.80**	**0**

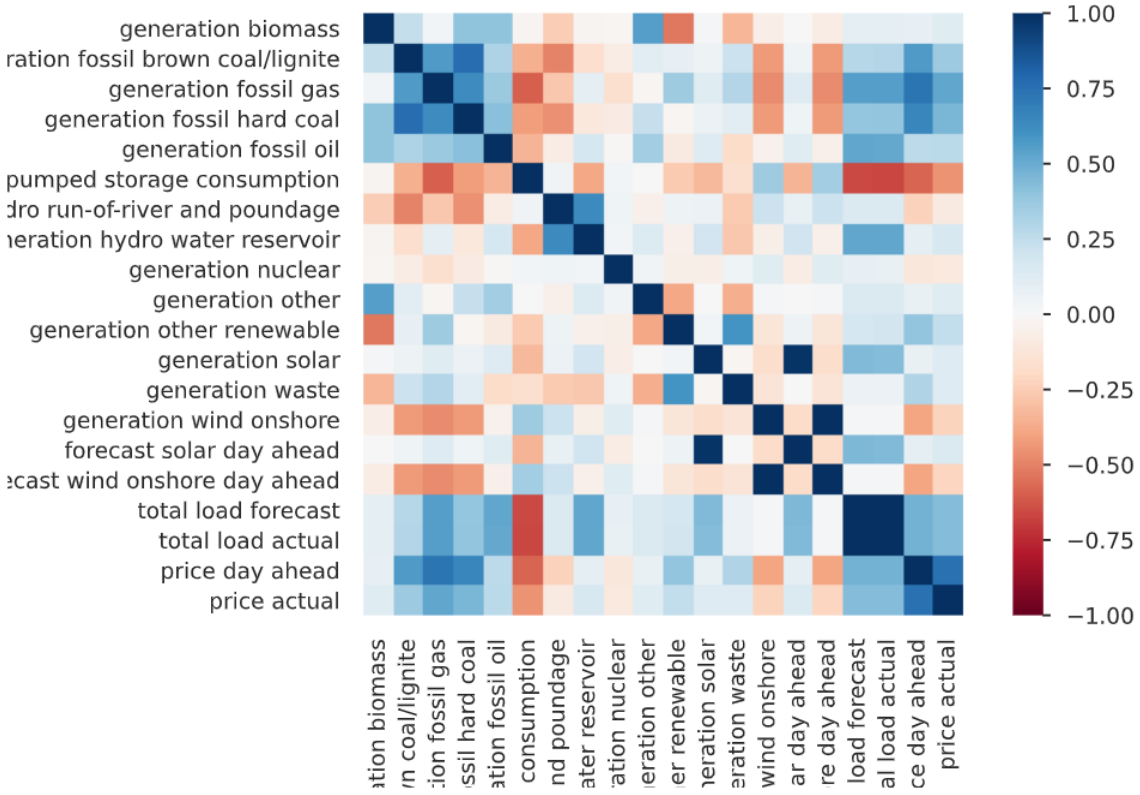


Figure 4: Correlation Matrix

5. APPROACH

1. **Choosing Dataset and theme of Project**
2. **Cleaning/Preparing Data**
    a. **Decide Dependent variable(s) for Regression and classification**
    b. **Finalize one dependent variable for each later**
    c. **First See what already done on dataset**
    d. **Then use RQ's keyword in general in google scholar to find relevant articles**
3. **Initial Problem analysis**
    a. **Write Literature review**
    b. **Focus on Research question not the ML**
4. **EDA**
    a. **Describe the data**
    b. **Any missing values, outlier**
    c. **Attribute type**
    d. **Descriptive Statistics(mean, std dev, min, max) of numeric attributes**
    e. **Box Plot of numeric attributes to determine if attributes are normally distributed**
    f. **Correlation Matrix to determine which attributes are related to each other**
5. **Feature Selection**
6. **Experimental Design and Cross Validation**
    a. **Regression (choose 3 algorithm) and Classification (choose 3 algorithm)**
    b. **Why choosing Regression and Classification**
        i. **If dependent variable is numeric then regr, if dep var is categorical then classification**
7. **Predictive Modelling**
    a. **(MAPE, RMSE for Regression and Accuracy, Precision, Recall for classification)**
    b. **Other Assumptions**
    c. **Performance evaluation**
8. **Conclusion and Recommedation**

6. REFERENCES

David Et Al. retrieved from https://arxiv.org/abs/1906.05433

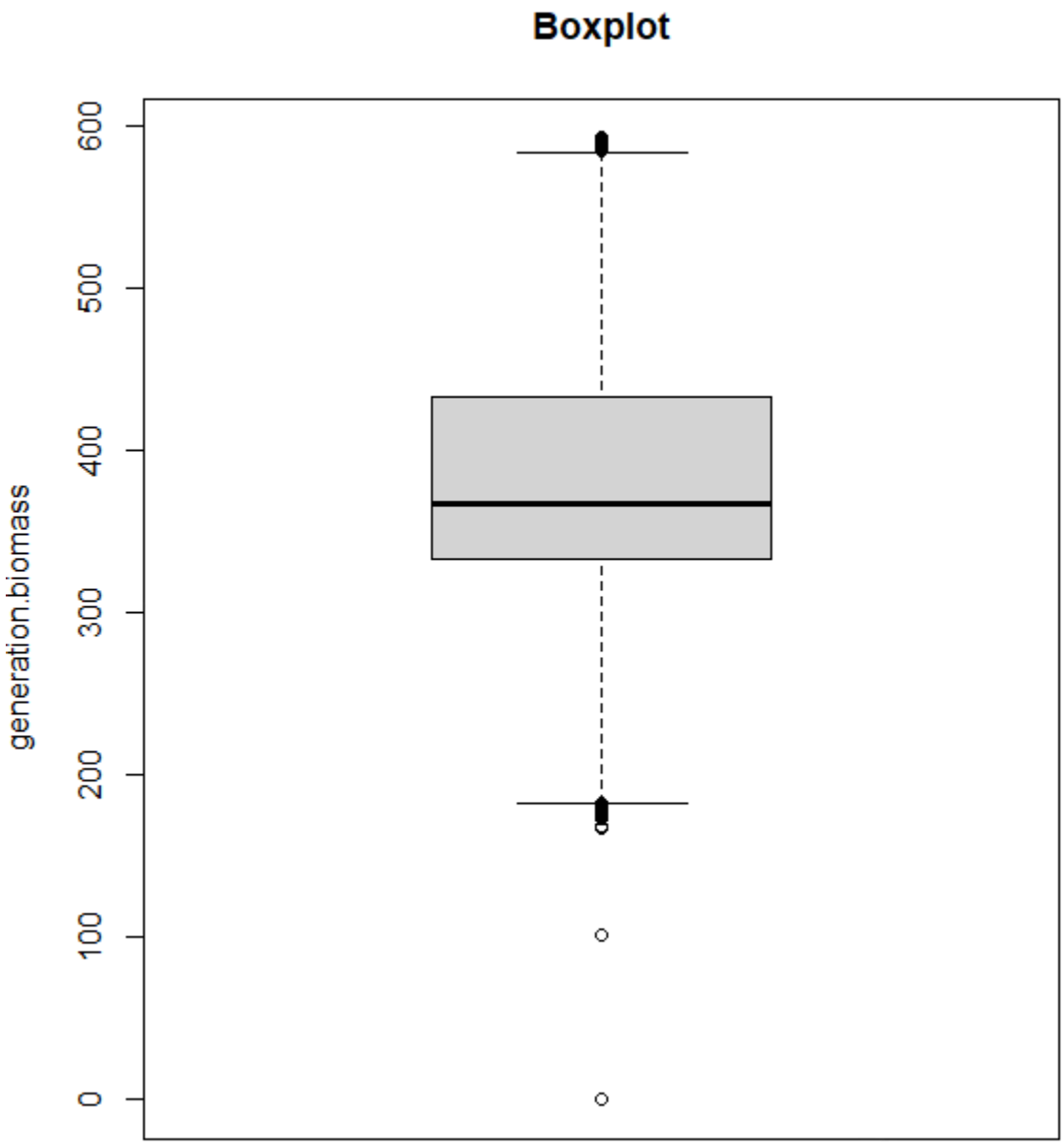
Kaggle retrieved from https://www.kaggle.com/datasets/nicholasjhana/energy-consumption-generation-prices-and-weather

Entsoe retrieved from https://transparency.entsoe.eu/dashboard/show

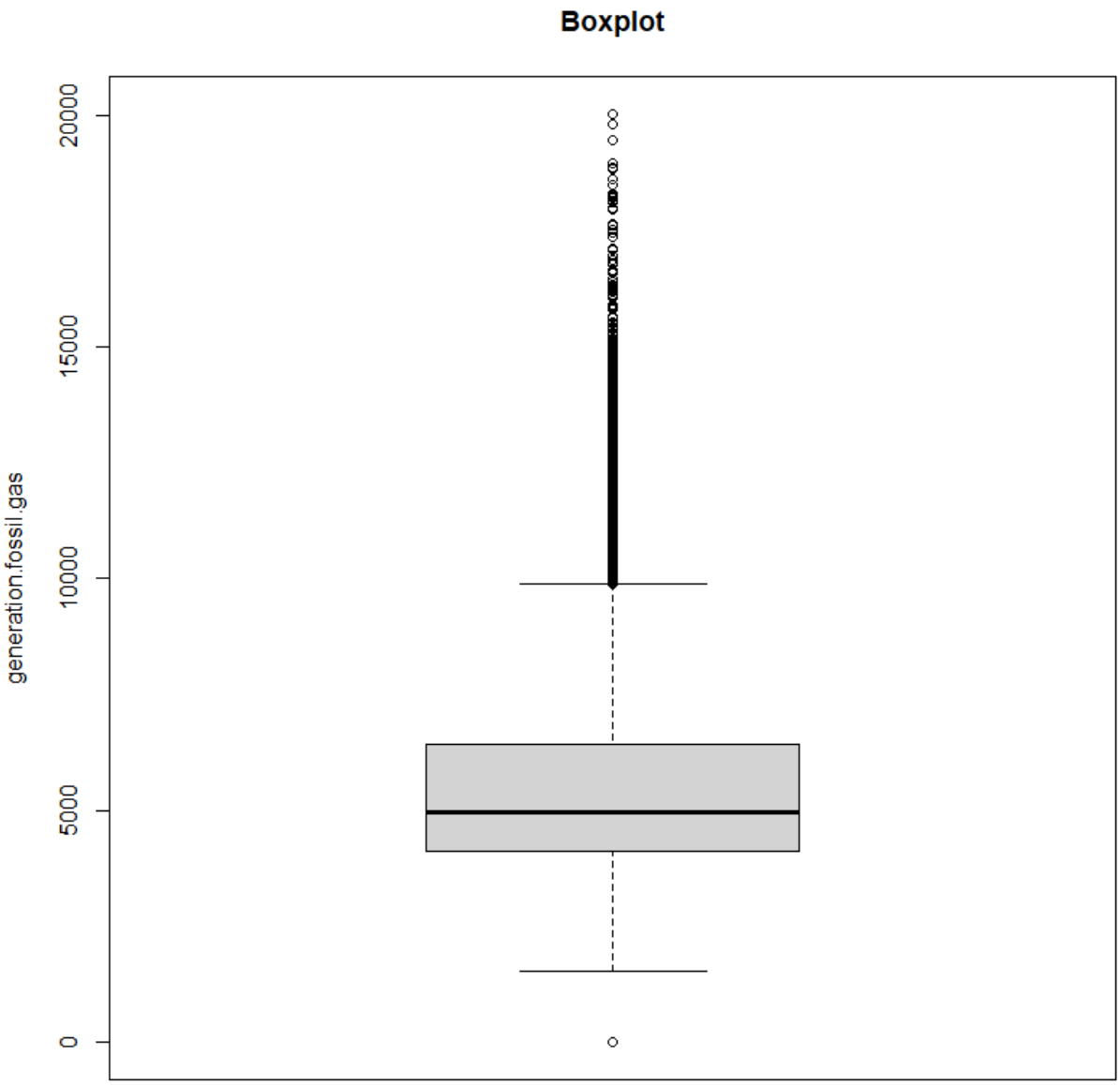
Esios retreived from https://www.esios.ree.es/en/market-and-prices?date=19-05-2023

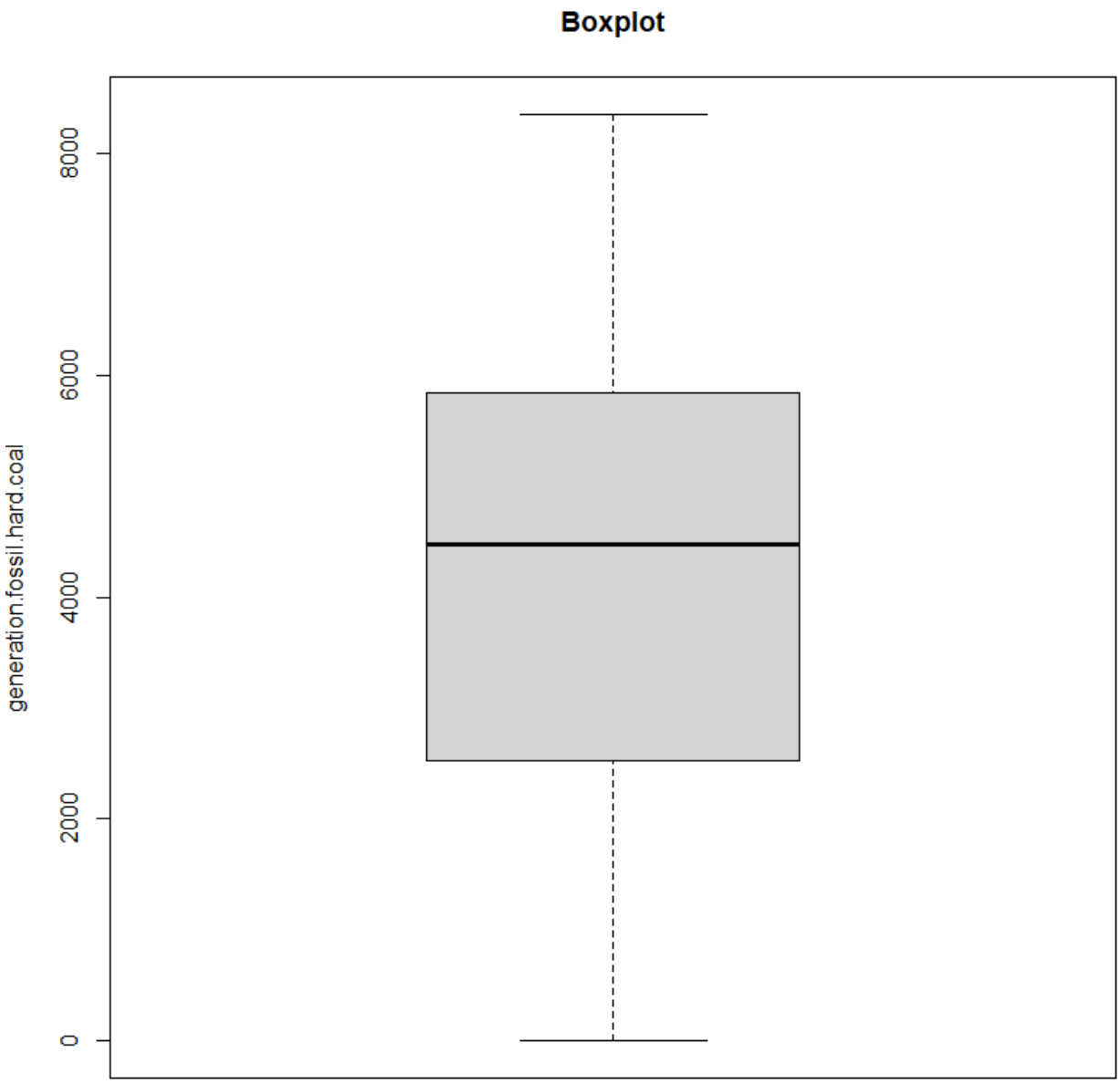
Openweather retrieved from https://openweathermap.org/api

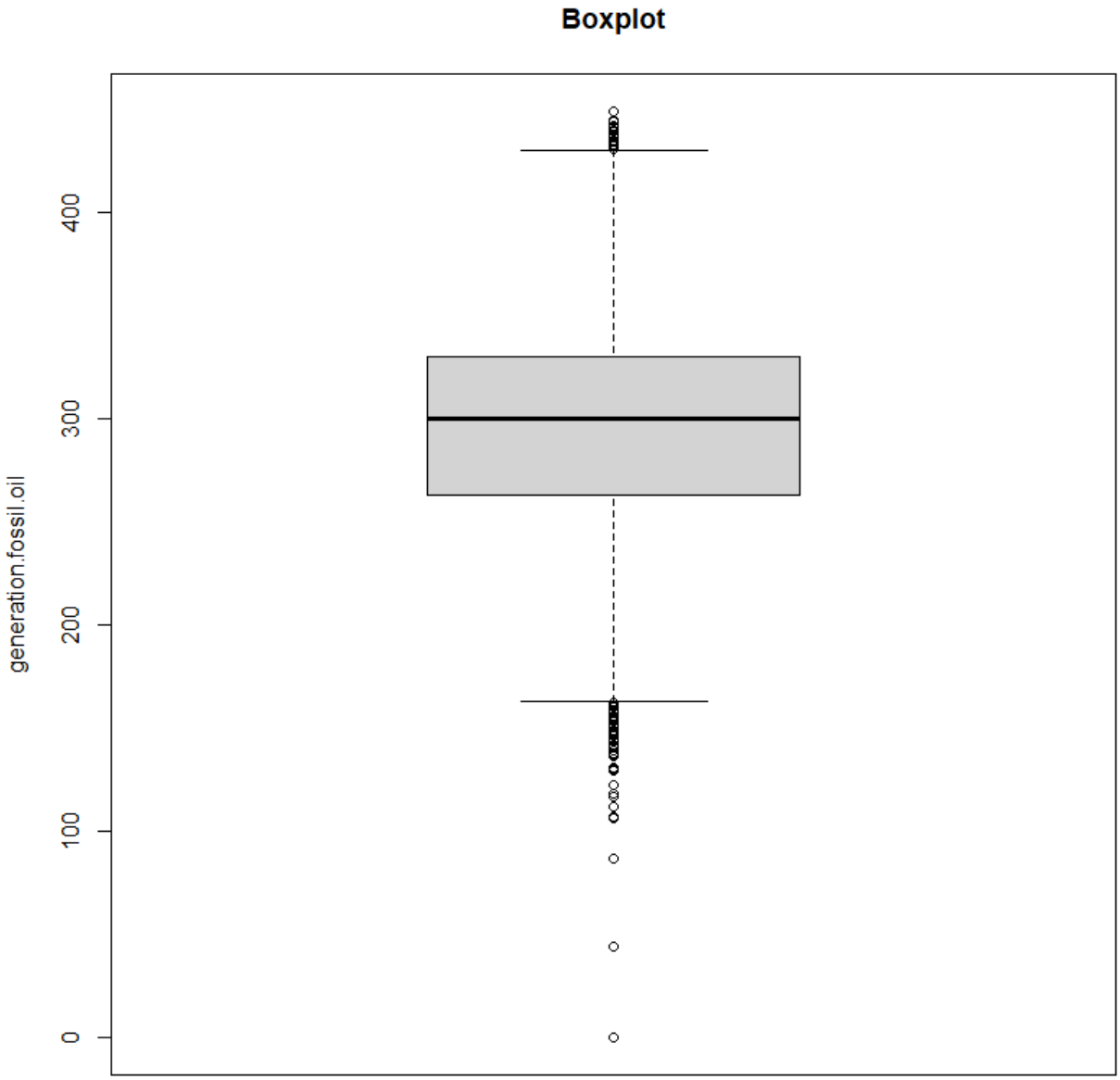
7. APPENDIX

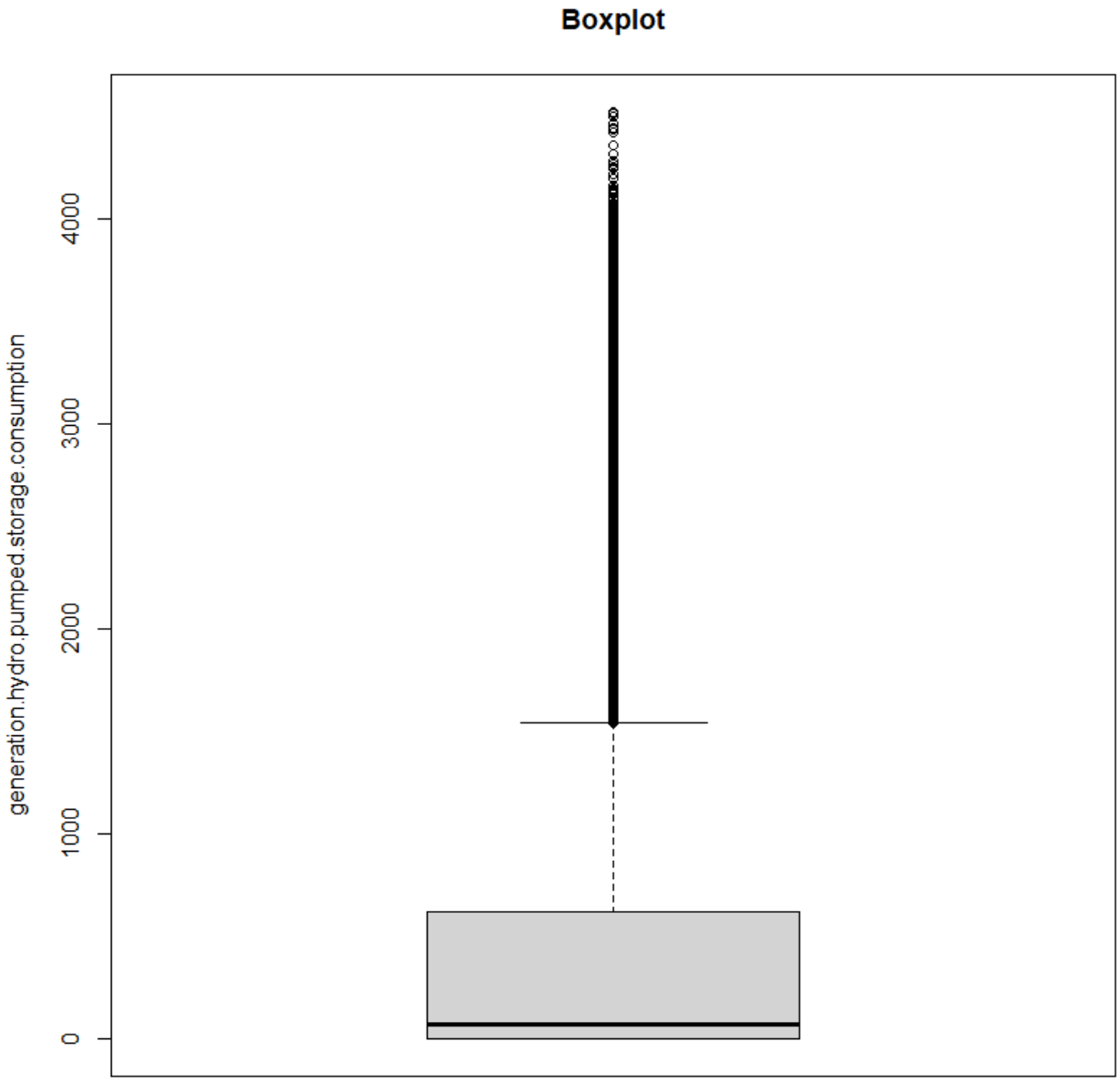


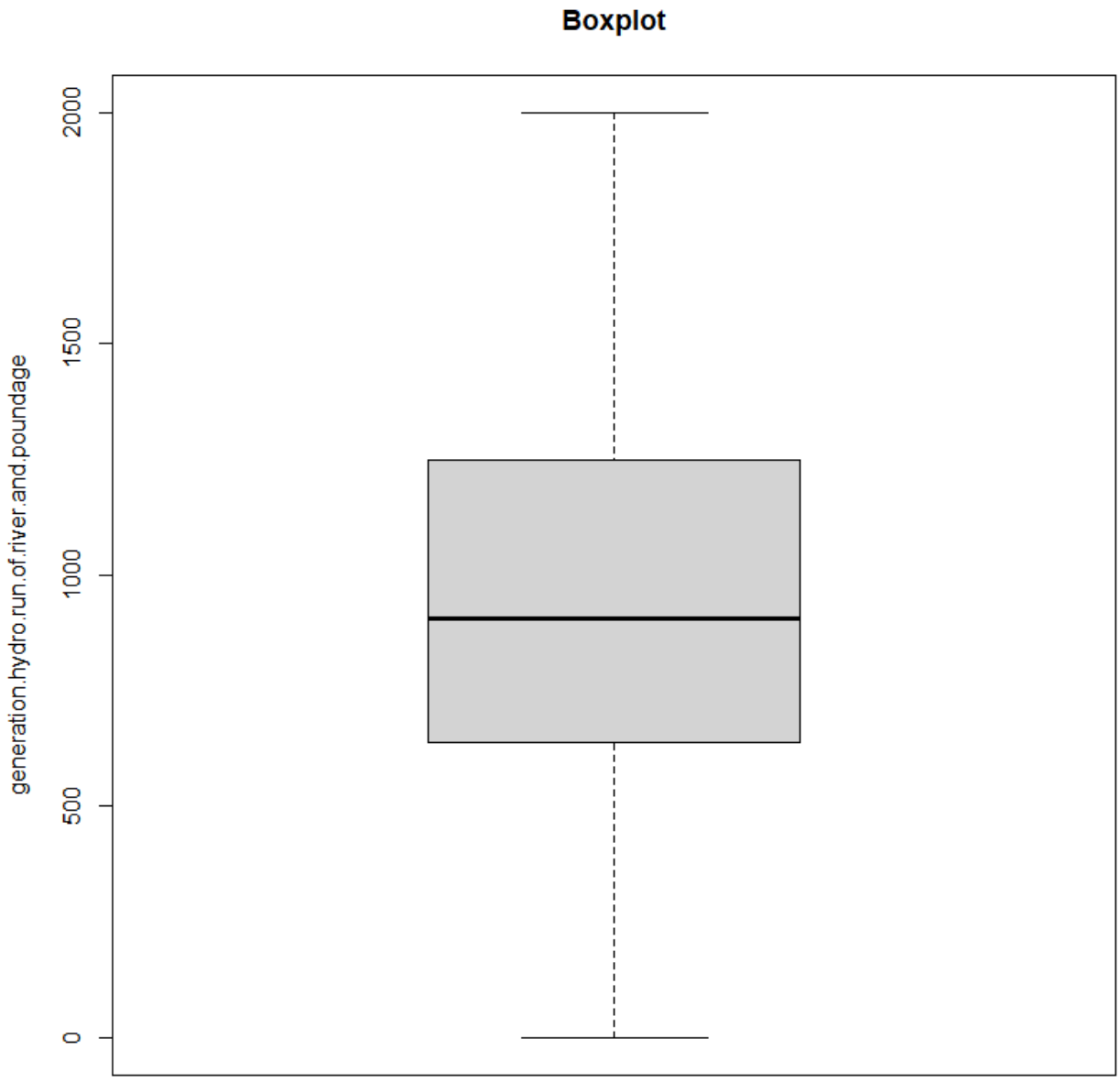

Boxplot
generation.fossil.brown.coal.lignite
0
200
400
600
800
1000

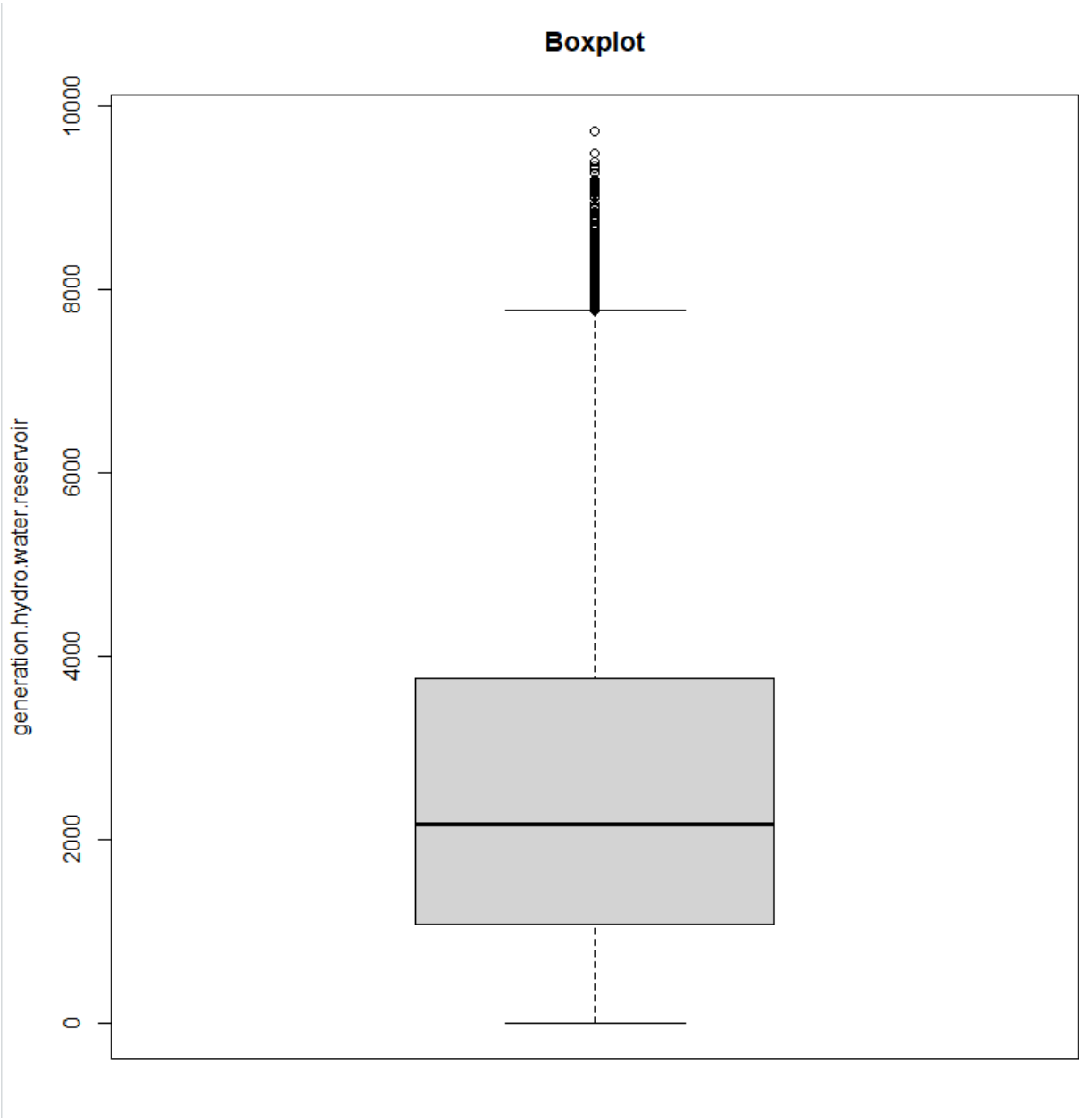

Boxplot
generation.fossil.gas
0
5000
10000
15000
20000

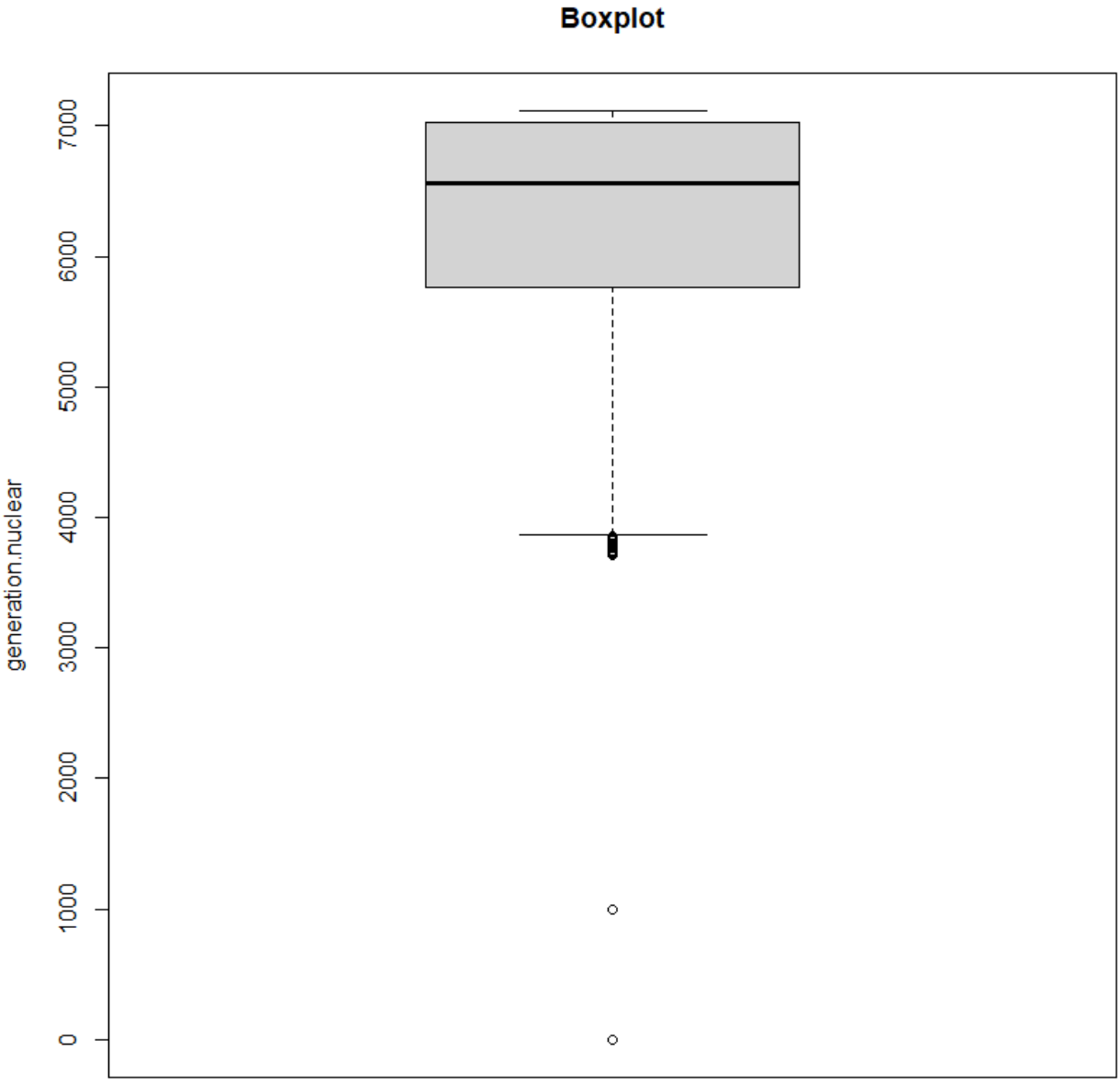

Boxplot
generation.fossil.hard.coal
0
2000
4000
6000
8000

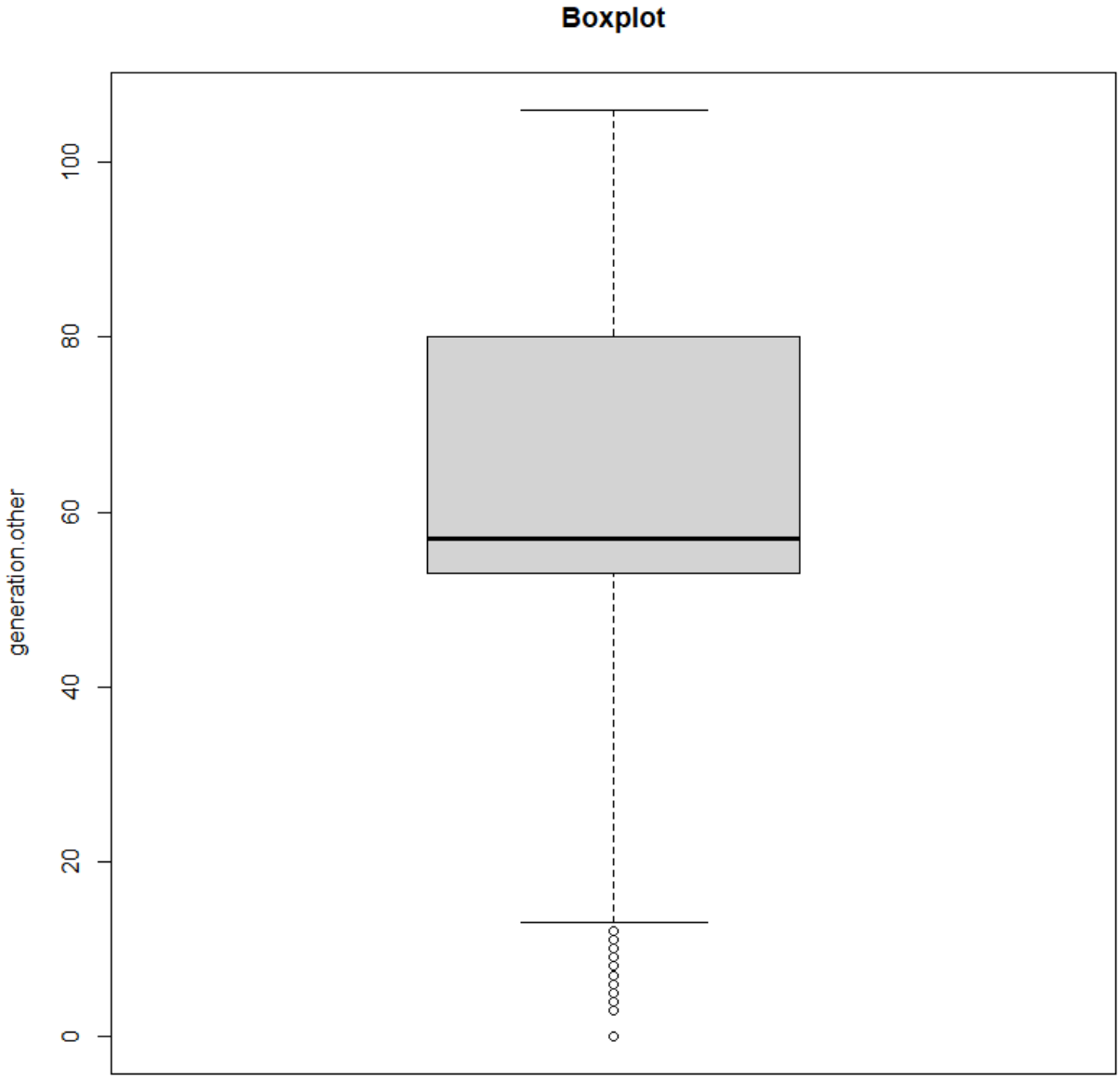

Boxplot
generation.fossil.oil
0
100
200
300
400

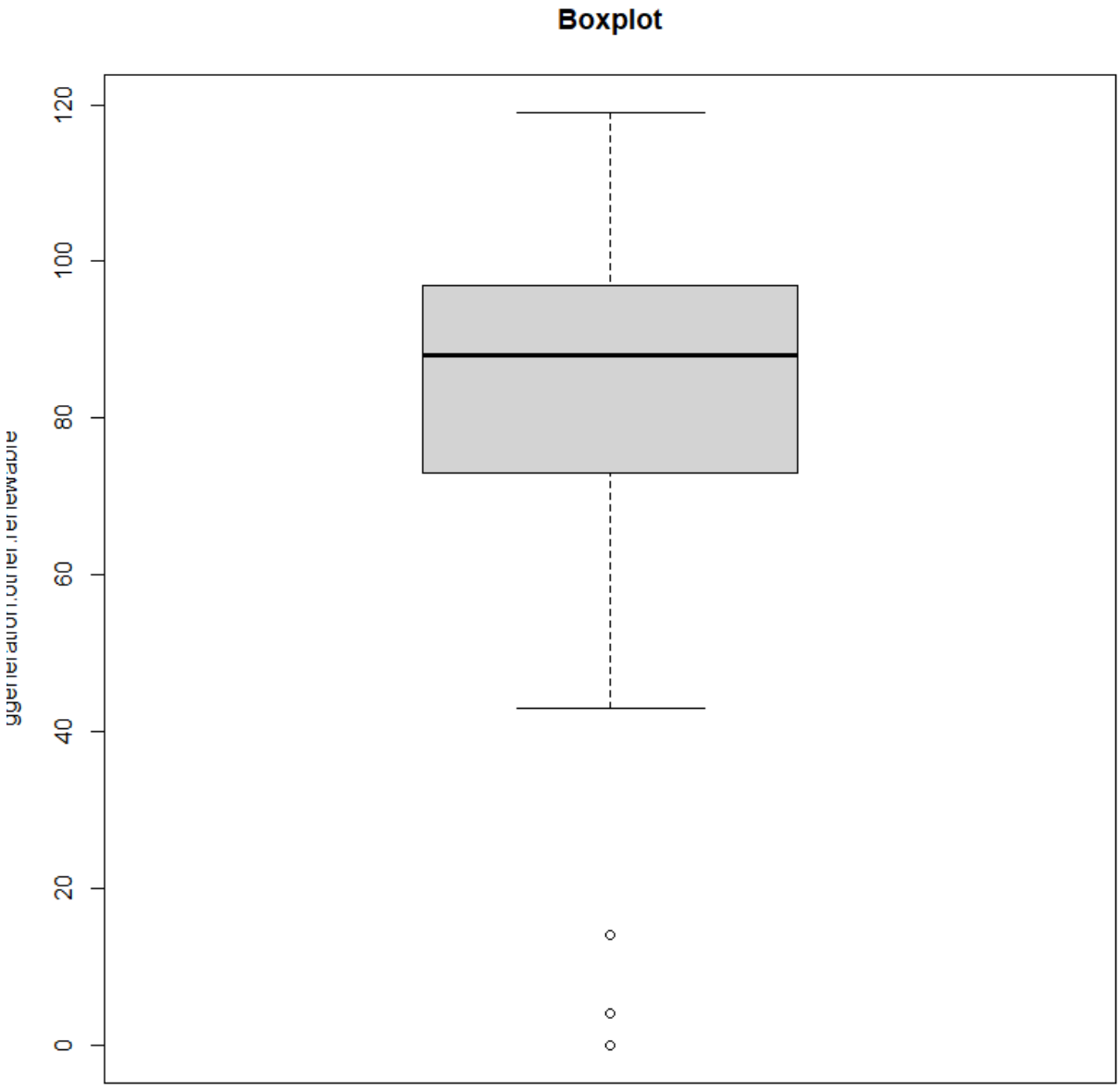

Boxplot
generation.hydro.pumped.storage.consumption
0
1000
2000
3000
4000

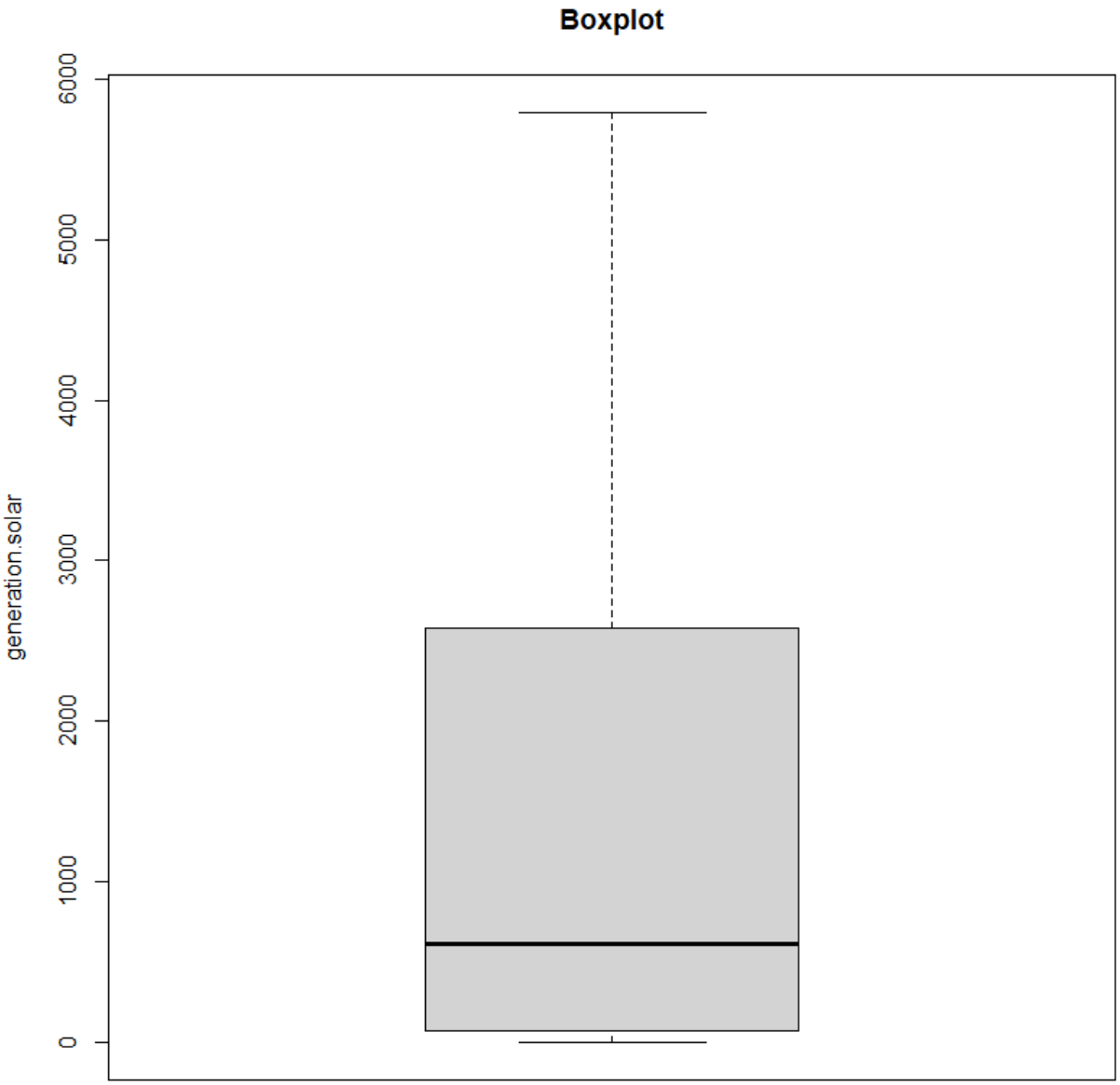

Boxplot
generation.hydro.run.of.river.and.poundage
0
500
1000
1500
2000

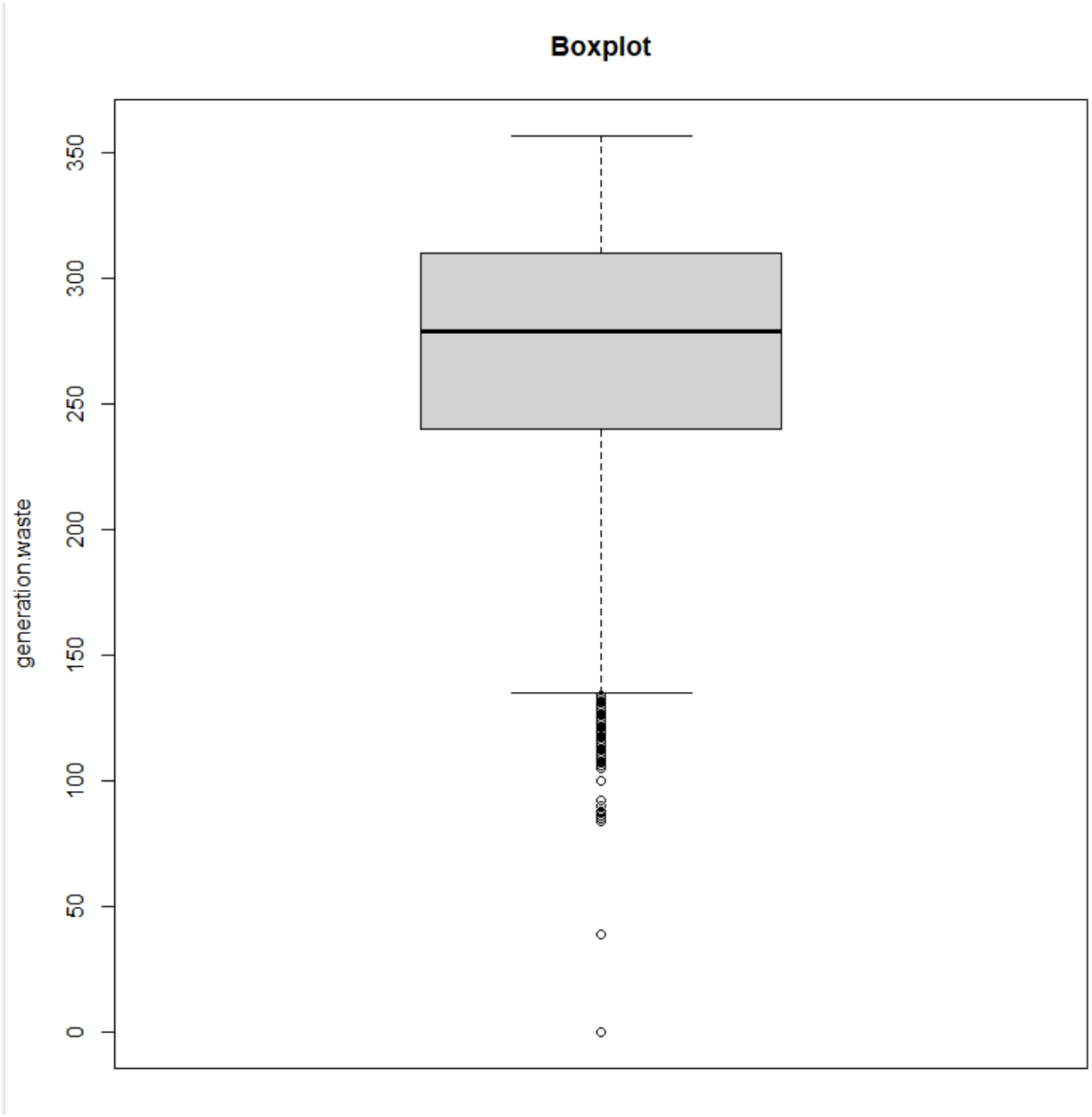

Boxplot
generation.hydro.water.reservoir
0
2000
4000
6000
8000
10000

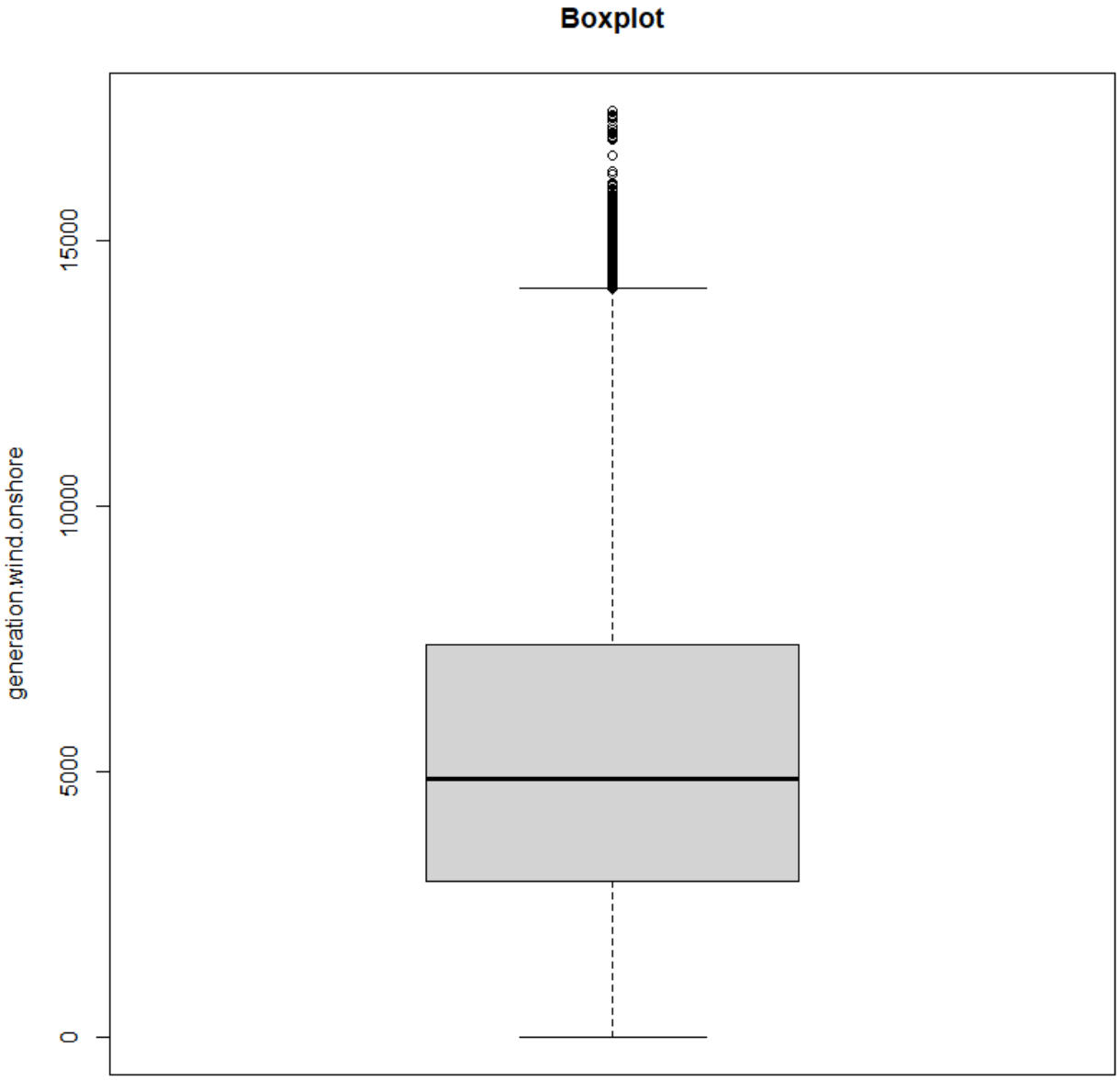

Boxplot
generation.nuclear
0
1000
2000
3000
4000
5000
6000
7000

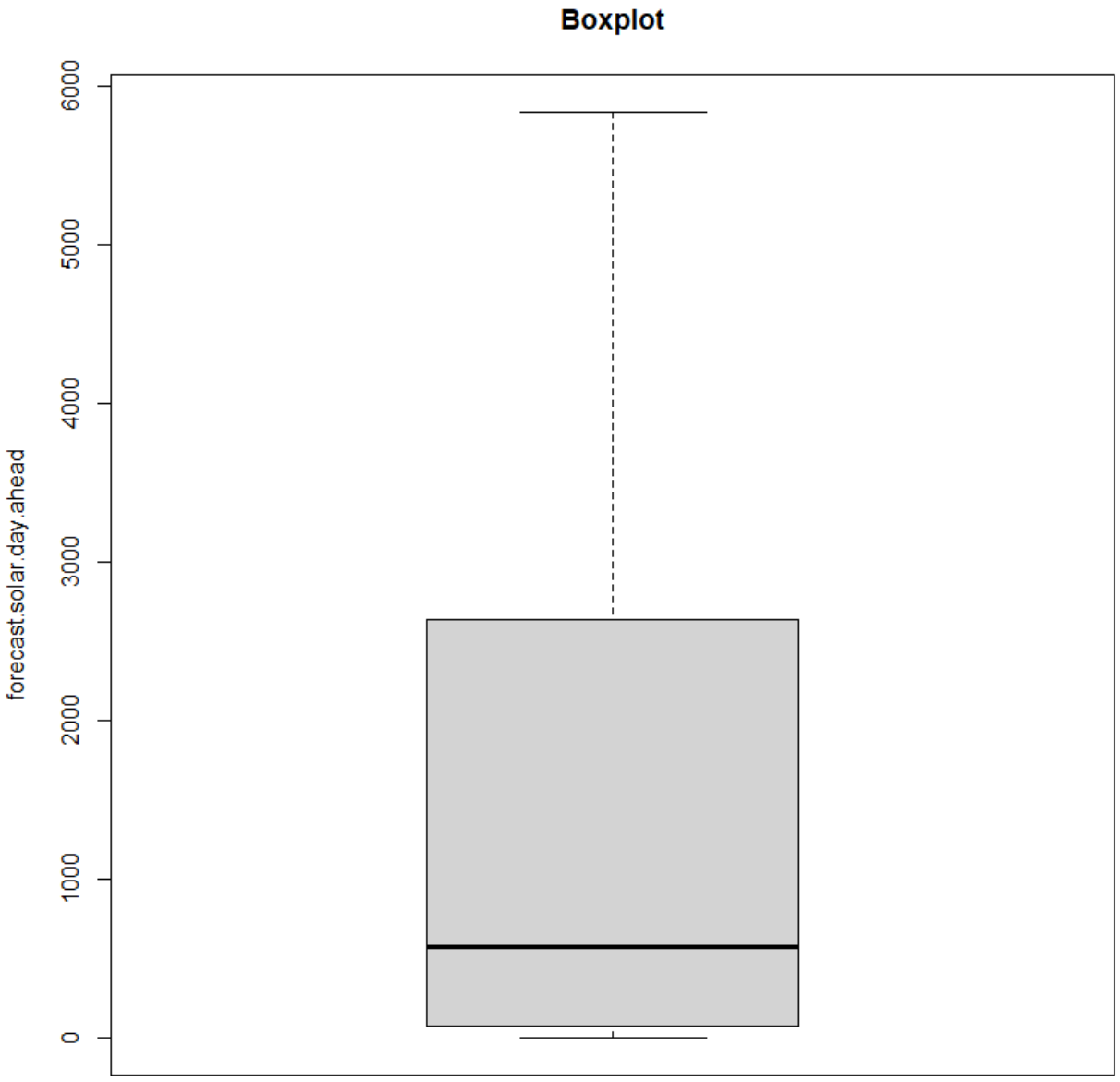

Boxplot
generation.other
0
20
40
60
80
100

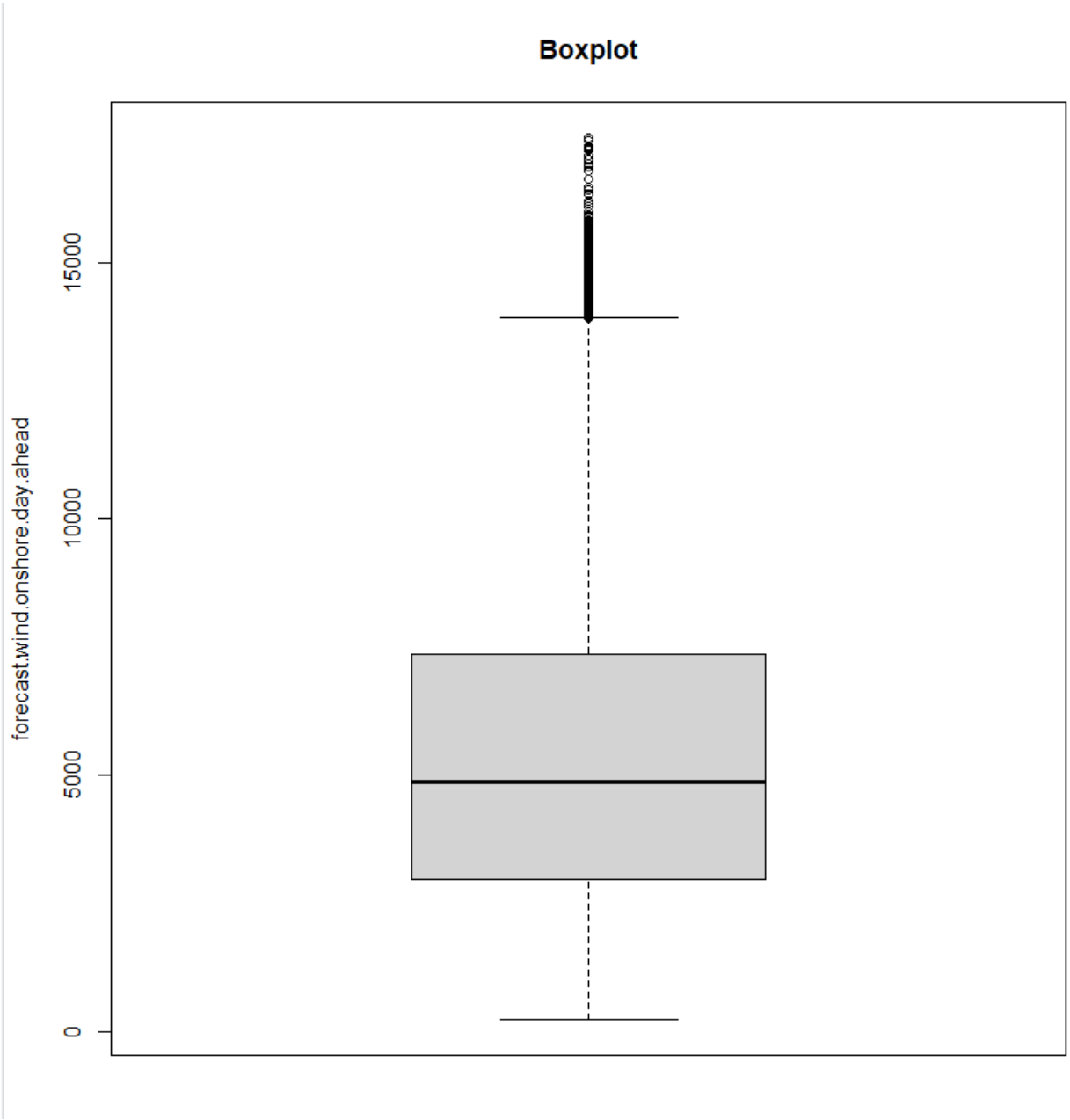

Boxplot
0
20
40
60
80
100
120

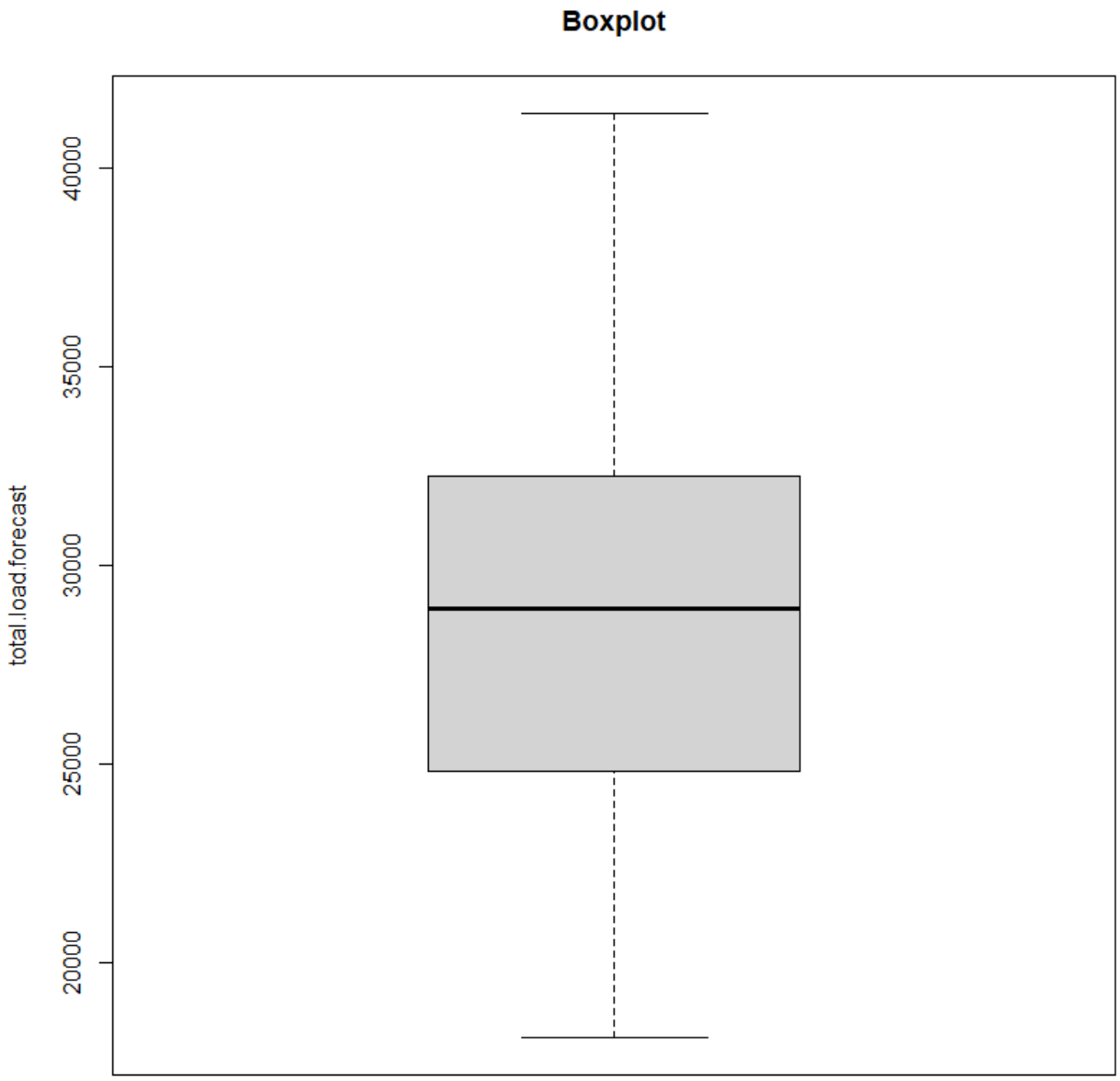

Boxplot
generation.solar
0
1000
2000
3000
4000
5000
6000

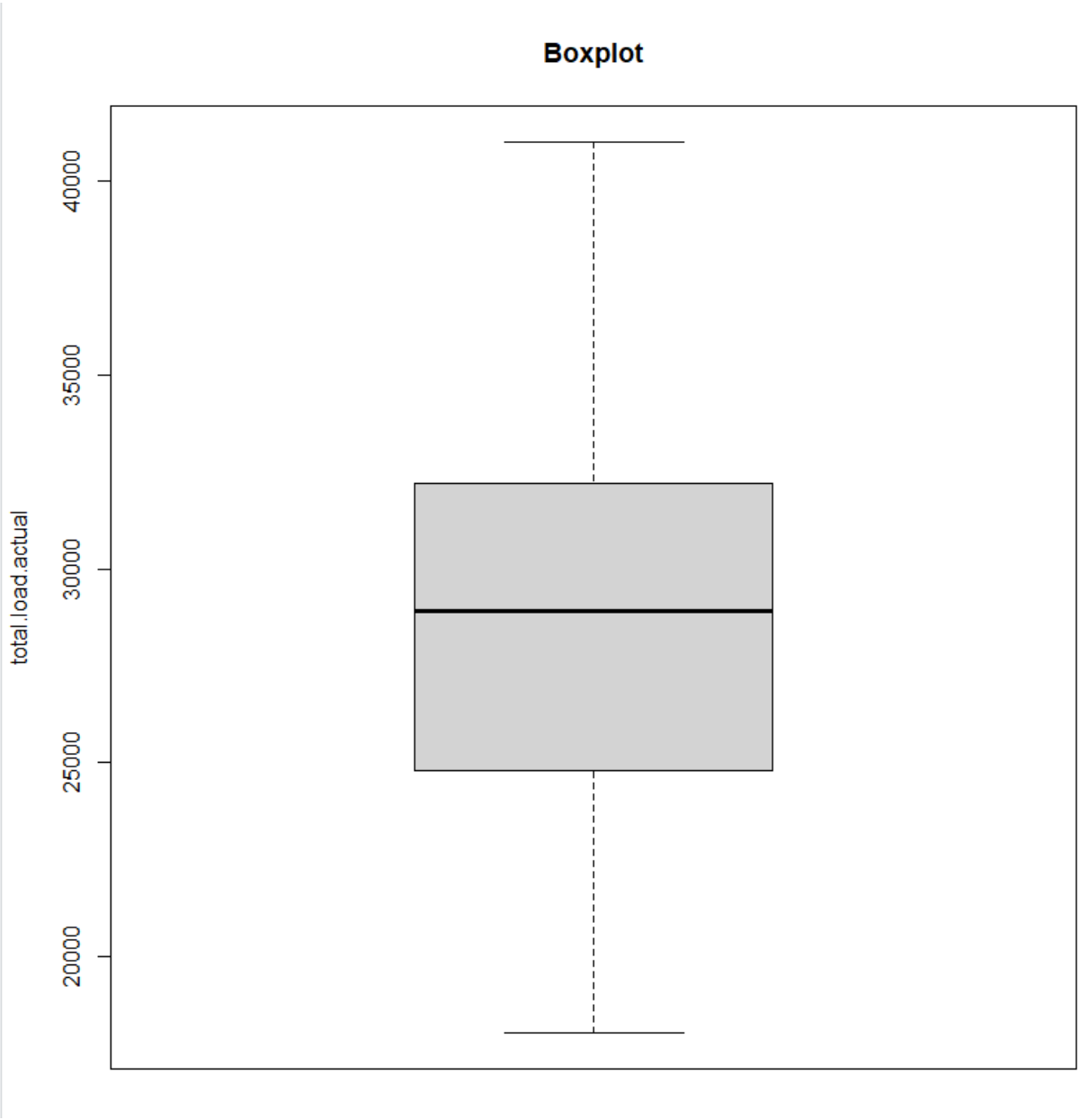

Boxplot
generation.waste
0
50
100
150
200
250
300
350

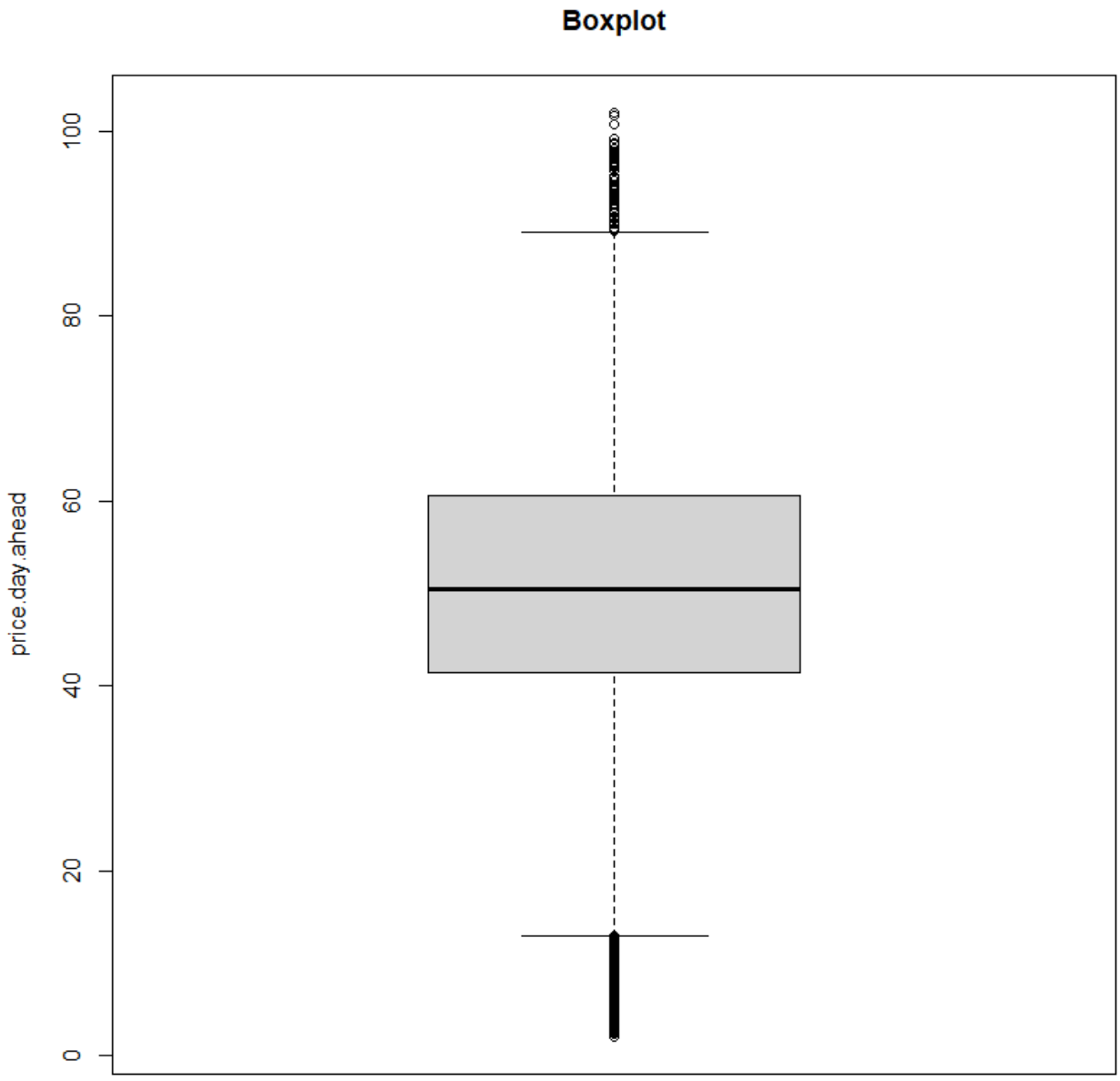

Boxplot
generation.wind.onshore
0
5000
10000
15000

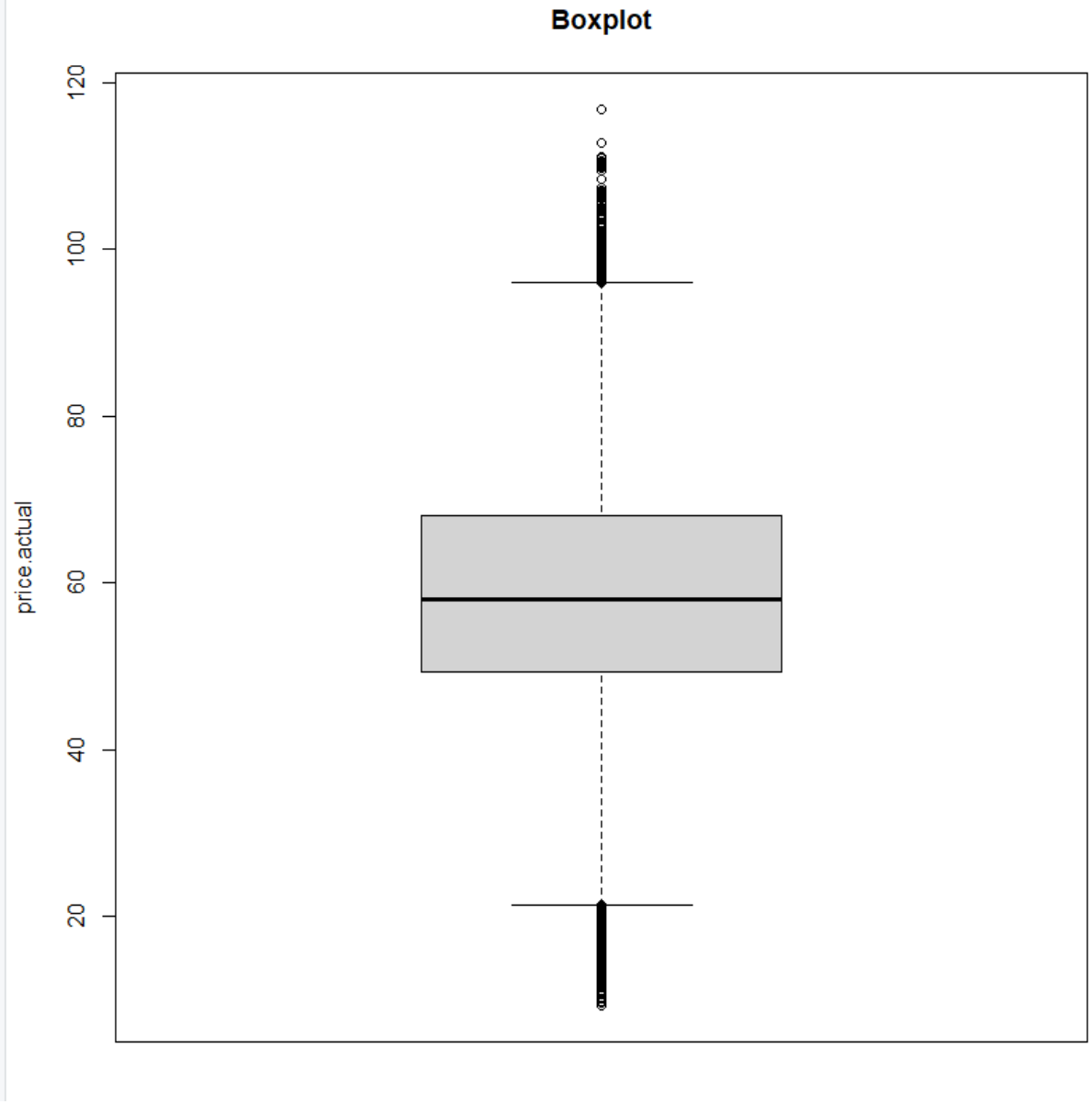

Boxplot
forecast.solar.day.ahead
0
1000
2000
3000
4000
5000
6000


Boxplot
forecast.wind.onshore.day.ahead
0
5000
10000
15000


Boxplot
total.load.forecast
40000
35000
30000
25000
20000


Boxplot
total.load.actual
20000
25000
30000
35000
40000


Boxplot
price.day.ahead
0
20
40
60
80
100


Boxplot
price.actual
120
100
80
60
40
20