<Title> **Predicting Energy Demand in Spain**

<Course> **CIND 820: Big Data Analytics Project**

<Section> **DAH**

<Student Name> **Muhammad Zaka Shaheryar**

<Student Number> **500648718**

<Supervisor Name > **Ceni Babaoglu**

<Date of Submission> **July 29th, 2023**

Ryerson University

Predicting Energy Demand in Spain

Table of Contents

Predicting Energy Demand in Spain

Acknowledgement

I want to convey thanks to Dr. Ceni Babaoglu for supervising me throughout the project. I also would like to thank Dr. Tamer Abdou and Riyad Hussain for their help throughout the project.

Finally, I would like to thank all my Professors and Teaching assistants, Program Director and Academic coordinator in my certificate program for teaching us valuable skills that we are able to apply in this project.

Predicting Energy Demand in Spain

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 has 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 has hourly data for electrical consumption and respective forecast by Transmission Service Operator (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 revised 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, how to accurately forecast daily energy demand?

The tools used are Python, Weka, Tableau, R, Excel. The systematic data analysis process approach is 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.

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 Linear 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

Predicting Energy Demand in Spain

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 burning 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 Github link for the project is provided below https://github.com/Zaka123456/CIND-820

Table 1: Literature Review Research:

No.	Research Question (RQ)
Q1	**Which regression technique will accurately forecast the daily energy consumption demand using hourly period?**
Q2	**How to accurately forecast energy demand 24 hour in advance compared to TSO?**
Q3	**Using classification, determine what weather measurement and cities influence most the electric demand, prices, and generation capacity?**
Revised Q3	**Using classification, how to accurately forecast daily energy demand?**

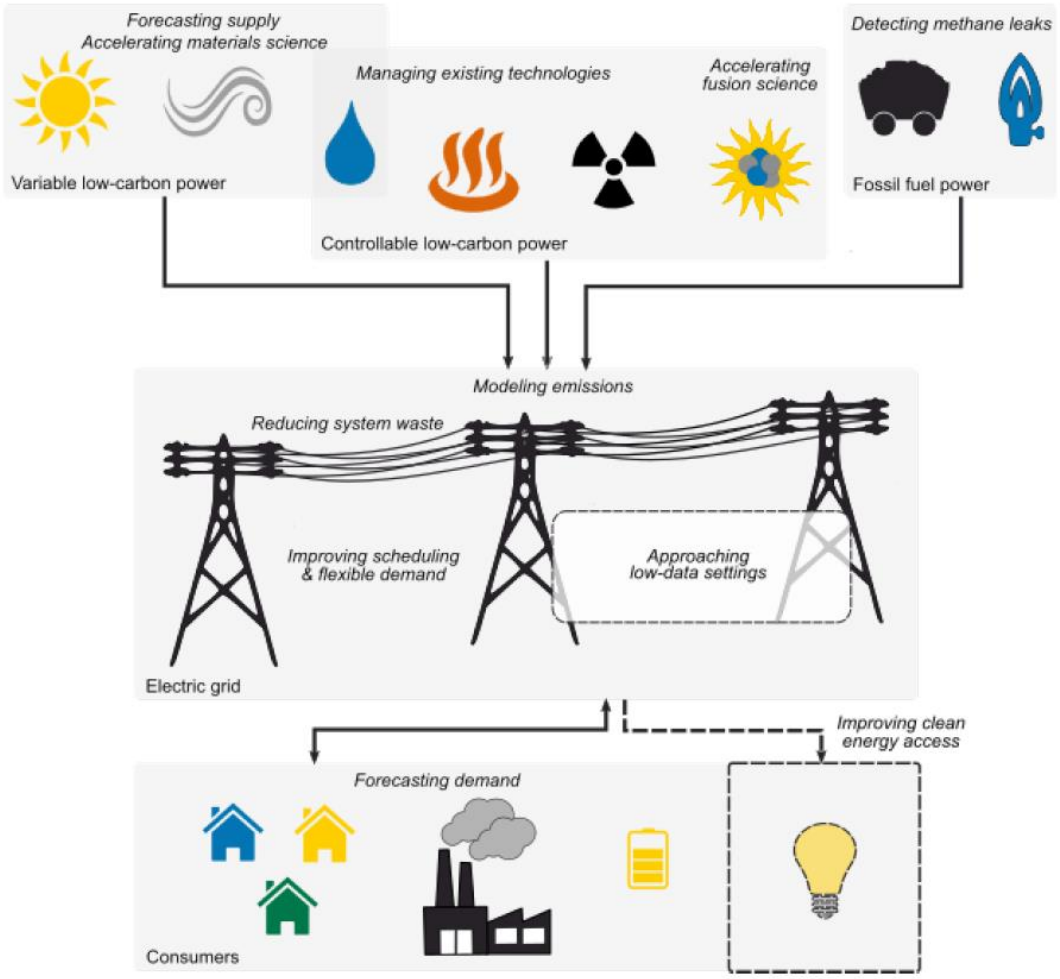


Figure 1: Selected opportunities to reduce GHG emissions from electricity systems using machine learning.

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 nonlinearity 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 papers are reviewed and key take-aways and how it can be useful to this project is provided below.

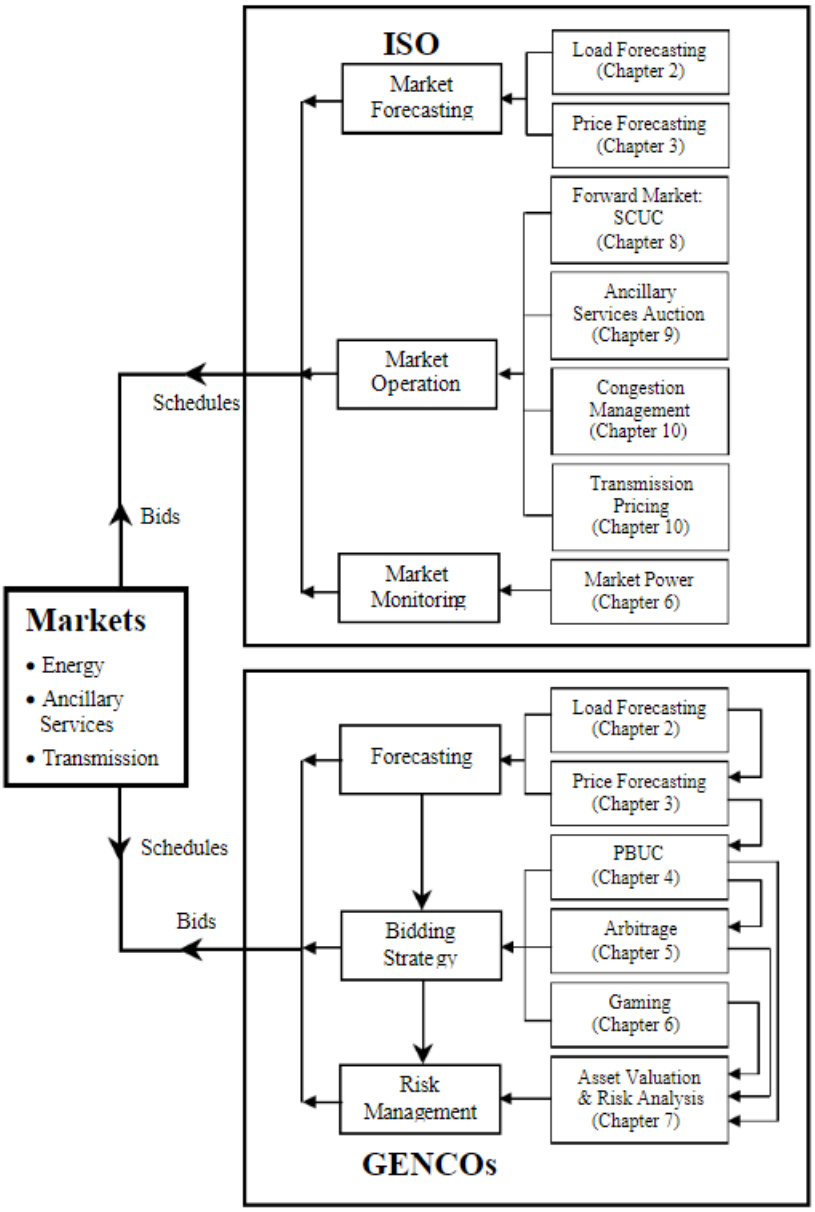


Figure 1.1 Restructured Electricity Market Operation

Figure 2: **Restructured electricity market operation**

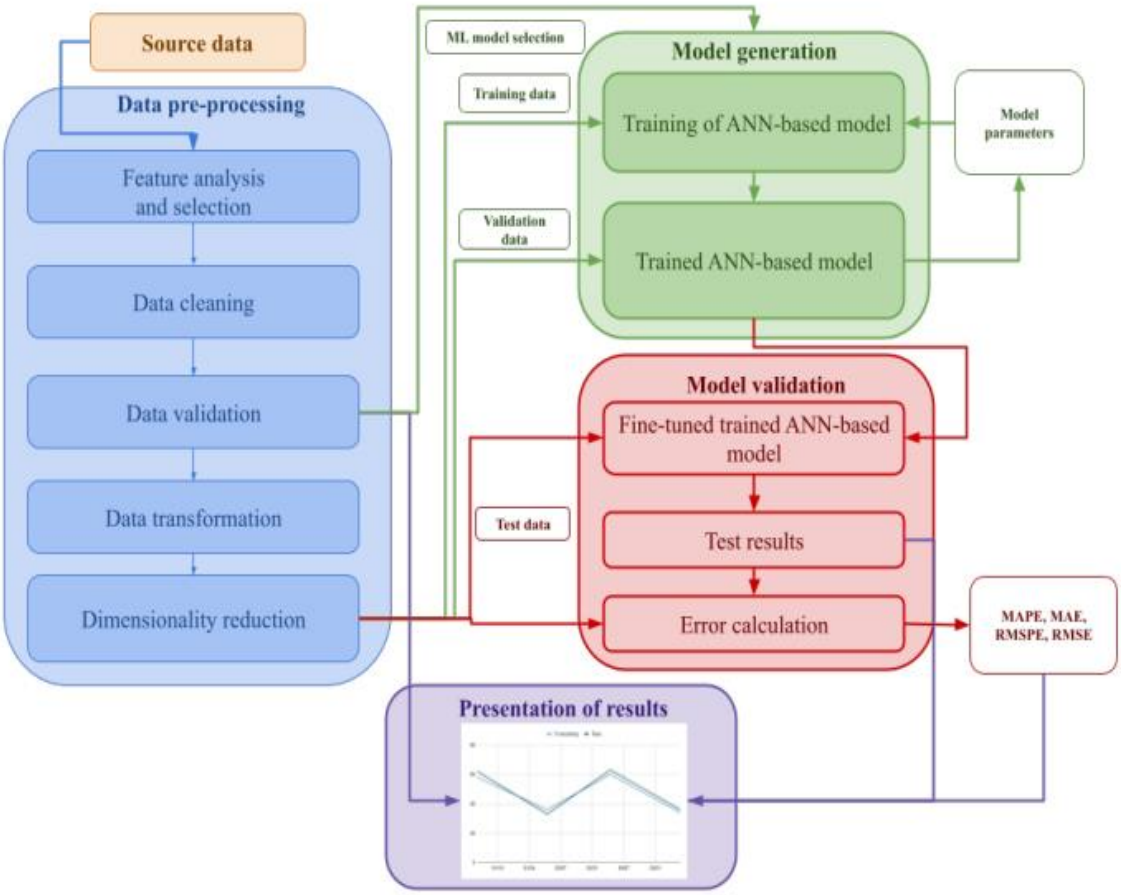


Figure 3: Generalization of steps documented in review papers

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 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 parameters are referred to as influence parameters as they influence the target parameter (load). Regression models and how to improve them is provided. Single vs multivariate regression models are 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 types of short-term electricity 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 Regulation Backpropagation is the best ML model. Furthermore, regression models performed well compared to advanced ML models. Moreover, almost all models 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 consists of three office buildings, and they all are managed by one 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 classifier, 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 and September) in year 2012 are used. Furthermore, three training methods are considered: Static, Accumulative and Sliding Windows. Cumulative and Sliding windows 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 allows 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 Networks (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-Neighbors 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 show that 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 and 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 simultaneously 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 provides useful insights regarding relations between input variables and three different loads and among three different loads as well. I can also apply the visualization like figure 4 in study as it gives 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. Coefficient of determine (R^2) is a performance metric. The results show 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 26^{th,} 2015. Seven regression models are used in study namely: Extra tress Regressor, AdaBoost, SVR, K-Neighbors 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 efficiency.

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%, relationship 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 like paper five above as it involves renewable energy source for power generation. Electricity renewables generates are not easily predictable and they vary 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 explores 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.

Predicting Energy Demand in Spain

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 hours 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.

Approach

- 1. Choosing Dataset and theme of Project
- 2. Cleaning/Preparing Data
    - Decide Dependent variable(s) for Regression and Classification.
    - Finalize one dependent variable for each later.
    - First See what already done on dataset.
    - Then use RQ's keyword in general in google scholar to find relevant articles.
- 3. Initial Problem analysis
    - Write Literature review.
    - Focus on Research question not the ML.
- 4. EDA
    - Describe the data.
    - Any missing values, outlier
    - Attribute type
    - Descriptive Statistics (mean, standard deviation, min, max) of numeric attributes
    - Box Plot of numeric attributes to determine if attributes are normally distributed.
    - Correlation Matrix to determine which attributes are related to each other.
- 5. Feature Selection
- 6. Experimental Design and Cross Validation
    - Regression (choose 3 algorithm) and Classification (choose 3 algorithm)
    - Why choosing Regression and Classification
        - If dependent variable is numeric then regression, if dependent variable is categorical then classification.

- 7. Predictive Modelling
    - (MAPE, RMSE for Regression and Accuracy, Precision, Recall for classification)
    - Other Assumptions
    - Performance evaluation
- 8.Conclusion and Recommendation
- 9. Limitation, future direction

Dataset Description

"Hourly energy demand generation and weather" (Kaggle) is the chosen dataset. Dataset is taken from Kaggle, and it contains two big csv files containing data of four years (2015-2018). The dataset contains data on electrical consumption, generation, pricing, and weather data for five cities in Spain. The data is retrieved from (Entsoe, Esios and Openweather). The links are given in references.

First File (energy_dataset)

First file is "energy_dataset" which has 35064 rows and 29 columns. This csv file has hourly data for electrical consumption and respective forecast by Transmission Service Operator (TSO) such as Spanish esios Red Electric Espana (REE) for consumption and pricing. Electric power generation and consumption is measured in mega watts (MW) and prices is in Euros. The attributes type is given in table 3. Out of 29 attributes, 26 are numeric, 3 are non-numeric. The attributes that have 99% or more missing values or zeros are dropped as there is no information regarding these attributes. They are provided in table 4. Using R and python, descriptive Statistics for remaining 20 numeric attributes is provided in Table 5. 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 3: Attributes Type of first file

Attribute	**Type**	**Attribute**	**Type**
time	**Non-numeric**	generation.nuclear	**Numeric**
generation.biomass	**Numeric**	generation.other	**Numeric**
generation.fossil.brown.coal.lignite	**Numeric**	generation.other.renewable	**Numeric**
generation.fossil.coal.derived.gas	**Numeric**	generation.solar	**Numeric**

Attribute	Type	Attribute	Type
generation.fossil.gas	**Numeric**	generation.waste	**Numeric**
generation.fossil.hard.coal	**Numeric**	generation.wind.offshore	**Numeric**
generation.fossil.oil	**Numeric**	generation.wind.onshore	**Numeric**
generation.fossil.oil.shale	**Numeric**	forecast.solar.day.ahead	**Numeric**
generation.fossil.peat	**Numeric**	forecast.wind.offshore.eday.ahead	**Non-numeric**
generation.geothermal	**Numeric**	forecast.wind.onshore.day.ahead	**Numeric**
generation.hydro.pumped.storage.aggregated	**Non-numeric**	total.load.forecast	**Numeric**
generation.hydro.pumped.storage.consumption	**Numeric**	total.load.actual	**Numeric**
generation.hydro.run.of.river.and.poundage	**Numeric**	price.day.ahead	**Numeric**
generation.hydro.water.reservoir	**Numeric**	price.actual	**Numeric**
price.actual	**Numeric**		

Table 4: Attributes with 99% or more zeros or missing values

generation.fossil.coal.derived.gas
generation.fossil.oil.shale
generation.fossil.peat

Table 4 continued
generation.geothermal
generation.marine
generation.wind.offshore
generation.hydro.pumped.storage.aggregated
forecast.wind.offshore.eday.ahead

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

No.	**Attribute**	**Mean**	**Standard deviation**	**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**

Table 5 continued	**Attribute**	**Mean**	**Standard deviation**	**Min**	**Max**	**NA**
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**
17	total.load.forecast	**28712**	**4594.101**	**18105**	**41390**	**0**
18	total.load.actual	**28697**	**4575**	**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**

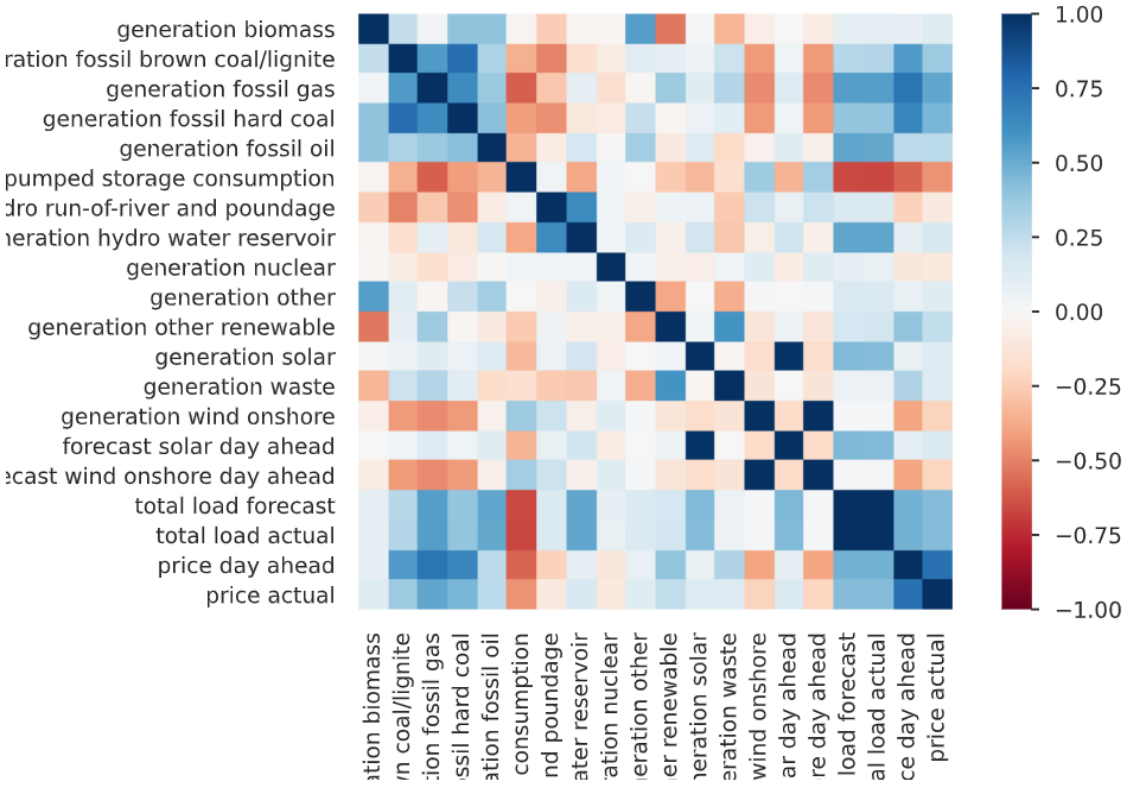


Figure 4: Correlation Matrix

Second file (weather_features)

The weather data contains hourly observations from five cities around Spain. Five Cities in Spain considered are Barcelona, Bilbao, Madrid, Seville, Valencia. This csv file has 178397 rows and 17 columns. temperature is measured in Kelvin, Pressure in hecto pascals, and wind speed in km/h. The attributes type is given in table 6. Out of 17 attributes, 12 are numeric, 5 are non-numeric. The attributes that have 99% or more missing values or zeros are dropped as there is no information regarding these attributes. There is only one attribute which is snow_3h so it is dropped. Also, redundant columns or categorical columns containing weather description are dropped as they donot influence the target. These are rain_3h, weather_id, weather_main, weather_descriptions, and weather_icon. Descriptive Statistics for remaining 9 numeric attributes is provided in Table 7.

Table 6: Attributes Type of second file

Attribute	Type	Attribute	Type
dt_iso	**Non-numeric**	Pressure	**Numeric**
city_name	**Non-numeric**	humidity	**Numeric**
temp	**Numeric**	wind_speed	**Numeric**
temp_min	**Numeric**	wind_deg	**Numeric**
temp_max	**Numeric**	rain_1h	**Numeric**
cloud_all	**Numeric**		

Table 7: Descriptive Statistic of Second file Numerical attribute with non-zero values

No.	Attribute	Mean	Standard deviation	Min	Max	NA
1	temp	**290**	**8**	**262**	**316**	**0**
2	temp_min	**288**	**8**	**262**	**315**	**0**
3	temp_max	**291**	**9**	**262**	**321**	**0**
4	pressure	**1073**	**6143**	**0**	**1008371**	**0**
5	humidity	**68**	**22**	**0**	**100**	**0**

Table 7 conti-ued	**Attribute**	**Mean**	**Standard deviation**	**Min**	**Max**	**NA**
6	wind_speed	**2.5**	**2**	**0**	**133**	**0**
7	wind_deg	**167**	**117**	**0**	**360**	**0**
8	rain_1h	**0.08**	**0.4**	**0**	**12**	**0**
9	cloud_all	**25**	**31**	**0**	**100**	**0**

Data Prepration

For regression analysis, **total.load.forecast** is selected as dependent variable initially. Since we must compare TSO forecast with forecast of machine learning model used. Therefore **total.load.actual** is selected as dependant variable. For classification analysis, **total.load.actual** is again selected as dependent variable after converting it to categorial feature namely **high_load_level** with two class true or false based on threshold value of 28700 MW since mean value is 28699.

After necessary packages imported are imported both files are uploaded in Google Colab. To overcome the daylight-saving issue in Madrid time zone, time field is converted to datetime index for both files. Co-ordinated Universal time (UTC) format is used for this purpose. Now both files as index set as datetime. There is only one hour offset between UTC and Madrid time which is ignored for analysis. All time values are selected from 2015 and above.

Then the columns mentioned in table 4 and redundant column of second file are dropped from their respective data frames. Then check for duplicates is carried out and if there are duplicates, they are removed from both files. There are no duplicates in first file. Then missing values are checked are if there is any they are imputed iteratively. File 2 does not have any missing values. Some columns have outlier in file 1 but they are ignored as they are not highly deviated. Moreover, in file 2 pressure, and wind_speed has outlier that are highly deviated, so they are treated accordingly. Temperature columns in file 2 shown values in Kelvin which is converted to degree Celsius for ease of analysis.

To get the idea about how much power is generated from different energy sources, one column is added namely “total generation”. Top five sources that generated almost 97% of power are Nuclear, Gas, Wind onshore, Hard coal, Hydro water reservoir, Solar, and hydro run of river and poundage. Also, Box plots and other important graphs are visualized as shown in appendix. It is observed that day ahead price is always cheaper than actual price for monthly sampled data. Similarly, for monthly sampled data, load forecast is almost like actual load which depicts that TSO forecasting provide results with great accuracy. Temperature Profile Plot for all 5 cities shows that all cities follow the seasonal pattern. Seville has the highest temperature while Bilbao and Madrid have the lowest temperature. Also, almost all 5 cities accounts for 20% frequency i.e., number of observations for each cities are almost equal, so location bias is negligible. Hourly load profile depicts that highest energy consumed is between 9-12 in the morning while lowest energy consumed is between 1-3 at night. Furthermore, almost all weekdays have similar load profile, while weekends (Saturday and Sunday) have much less consumption as expected. Moreover, Sunday has the

least consumption while Wednesday has the highest. In addition to that, monthly load prolife shows that Winter months (January and February) and Summer months (June, July) have greatest electricity usage while Spring months (April and May) has lowest. The distribution check whether it is normal or not is checked using histogram, as shown in Appendix.

Since second file has 5 time more data then first file, so to make it easier to combine both files, first mean value is taken across all 5 cities which make second files almost like first file in term of number of records. In this way second file only have one value for each time index rather than five. Then both data frame files are combined on date index column. The number of rows and columns in combined file is now 34468 and 31 respectively.

Feature Selection

To select the relevant feature for model training, Correlation matrix is checked. Following table shows the correlation coefficient between input parameters and target (total load actual).

Table 8: Correlation with Target

Attribute	**Correlation Coefficient with Target**
pressure	0.01
humidity	**-0.37**
wind_speed	**0.20**
wind_deg	-0.09
rain_1h	**0.02**
snow_3h	-0.01
clouds_all	0.01
temp_C	0.20
temp_C_max	**0.20**
temp_C_min	0.21

Attribute	Correlation Coefficient with Target
generation biomass	**0.08**
generation fossil brown coal lignite	**0.28**
generation fossil gas	**0.55**
generation hard coal	**0.40**
generation fossil oil	**0.50**
generation hydro pumped storage consumption	-0.56
generation hydro run of river and poundage	**0.11**
generation hydro water reservoir	**0.48**
generation nuclear	**0.09**
generation other	0.10

Attribute	**Correlation Coefficient with Target**
generation other renewable	**0.18**
generation solar	0.39
generation waste	0.08
generation wind onshore	0.04
total generation	**0.81**
forecast solar day ahead	**0.40**
forecast wind onshore day ahead	0.04
total load forecast	1
price day ahead	0.47
price actual	**0.43**

Based on the above results, following features are selected for regression analysis (bolded in above table) based on their high correlation value with target or low correlation with target but they are needed for energy forecast such as weather variables like rain_1h.

- **total generation**
- **humidity**
- **wind_speed**
- **rain_1h**
- **temp_C_max**
- **generation biomass**
- **generation fossil brown coal/lignite**
- **generation fossil gas**
- **generation hard coal**
- **generation fossil oil**
- **generation hydro run of river and poundage**
- **generation hydro water reservoir**
- **generation nuclear**
- **generation other renewable**
- **forecast solar day ahead**
- **price actual**

Afterward, smaller data frame with important features mentioned above and target feature is utilized for modeling. Among the above selected features, some of them are not normal. They are forecast solar day ahead, generation fossil brown coal lignite, generation nuclear, rain_1h, and wind_speed after treatment.

Modeling

Now for Regression analysis, we have 16 predictors and one target. Target variable is **total load actual** measured in MW. Random 80:20 split is done for training and testing respectively. Therefore, 27574 (80%) of records are for training and rest are for testing. Based on research question 1 and 2, data is sampled on hourly basis and daily basis respectively. After resampling to daily basis, the number of rows reduced to 1460. Four Regression model are used namely, Linear Regression, KNN Regression, Decision tree Regression, and Random Forest Regression. The results are shown in Table 9 and 10. Research question 3 is revised and one categorical column "**high_load_level**" is created as mentioned before. This categorical column is now target variable for classification analysis. It has two classes, true or false based on whether the total load actual is greater than 28700 MW or not. If it is greater or equal to 28700 MW it is true otherwise false. There is almost no class imbalance as true is 49% and false is 51%. Then data is normalized and 80:20 train:test split is carried out. Here predictors are 17 now because total load actual is also predictor while target is high_load_level. Three classification models are used namely, Logistic regression classifier, Knn classifier, and Decision tree classifier. The result is shown in Table 11.

Results and Discussion

Table 9: Regression analysis for hourly period

Model	RMSE	R^2	Rank
Linear Regression	1748	0.85	**3rd**
KNN Regression	1430	0.90	**2nd**
Decision Tree Regression	1693	0.86	**4th**
Random Forest Regression	1155	0.94	**1st**

Therefore, for hourly period the best learned model is Random Forest Regression with R square of 0.94 which means it can explain 94% of variation in dependent variable due to independent variables. Also, RF regression has RMSE of 1155 which is lowest among all four models.

Table 10: Regression analysis for daily period

Model	MAPE	RMSE	R^2	Rank
TSO Forecast	0.008	3.09	-	**1st**
Linear Regression	2.93	1042	0.86	**2nd**
KNN Regression	3.49	1296	0.78	**4th**
Decision Tree Regression	3.81	1479	0.71	**5th**
Random Forest Regression	2.86	1058	0.85	**3rd**

Therefore, all four models are unable to beat TSO forecast since they all have MAPE > 0.0082 and RMSE > 3.09. The next best model came out to be Random Forest Regressor with MAPE of 2.86 and RMSE of 1058.

Table 11: Classification analysis for daily period

Model	Accuracy	Precision	Recall	Rank
Logistic Regression Classifier	93.5%	93.5%	93.5%	**2nd**
KNN Classifier	90%	90%	90%	**3rd**
Decision Tree Classifier	100%	100%	100%	**1st**

There seem to be some issue with Decision tree classifier because all three performance measures are 100%. This need to be explored further. The next steps are to refine these models and make them robust. Second best is Logistic Regression classifier which gives 93.5% accuracy. Also, Logistic regression perform better than KNN classifier in term of Precision and Recall. Therefore, Logistic Regression classifier is better in predicting positive class and has better true positive rate compared to KNN classifier.

Analysis limitations

There are many limitations as well as shortcomings in the analysis. There seem to be issue in Decision tree classifier due to which all three performance measures are 100% which is not realistic. Also for classification analysis, normalization is done before splitting the data for training and testing. Cross validation is missing. Normality is just assumed using histogram rather than performing statistical test. Outliers are not treated properly. Also, hourly and day of week load profile depict outlier which are not treated.

Conclusion and Recommendations

There is a demand shortage over four years which shows that forecasting can be useful in planning and dispatching resources. Also adding time feature increase accuracy slightly. Analysis limitations provides a ground for future work which can improve the model performance and make the results more realistic. Cross validation should be done together with principal component analysis. After this the Regression and classification models should be rebuilt and re-run which can improve their performance.

References

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

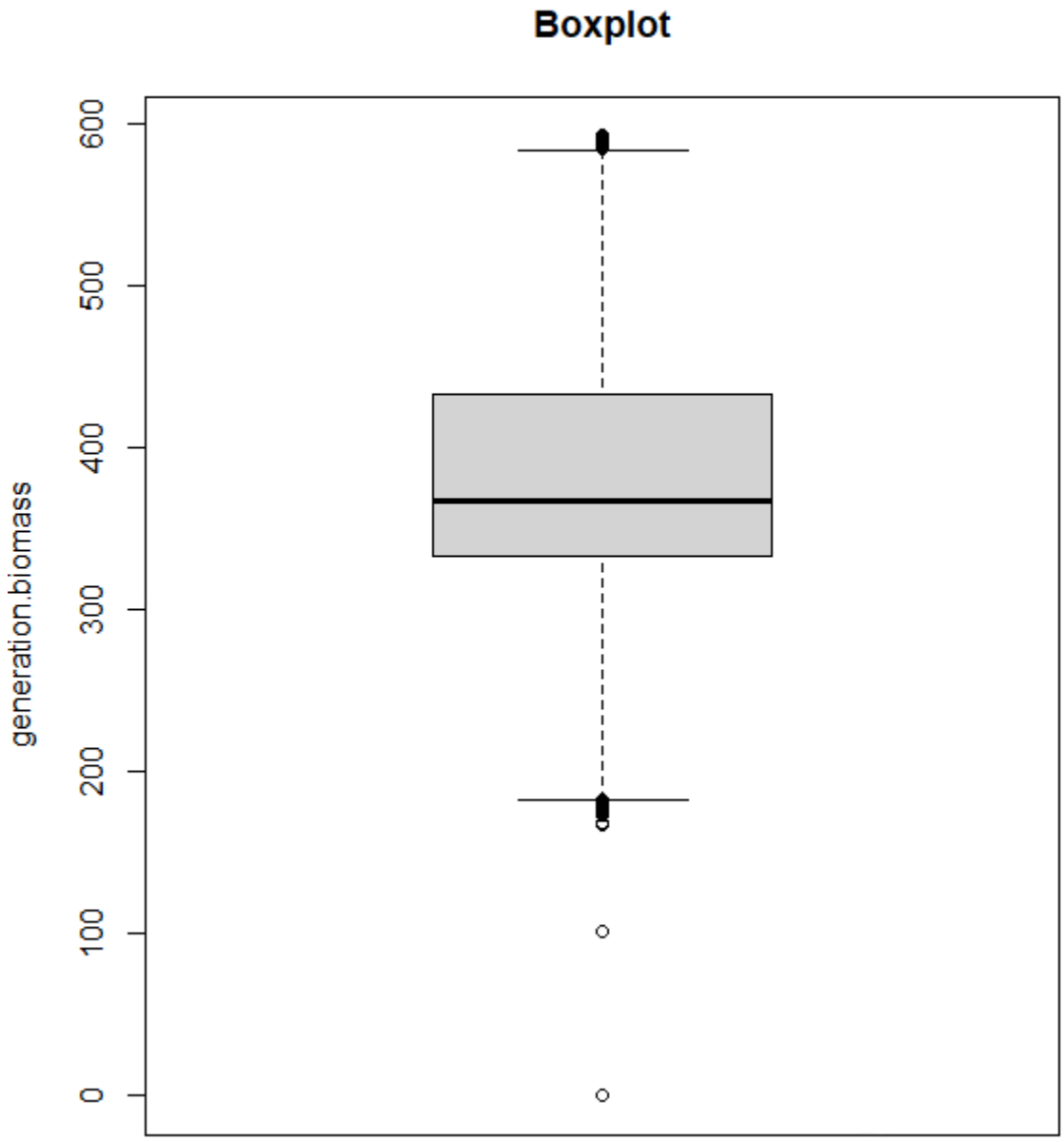
Kaggle retrieved from https://www.kaggle.com/datasets/nicholasjhana/energy-consumption-generation-prices-and-weather

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

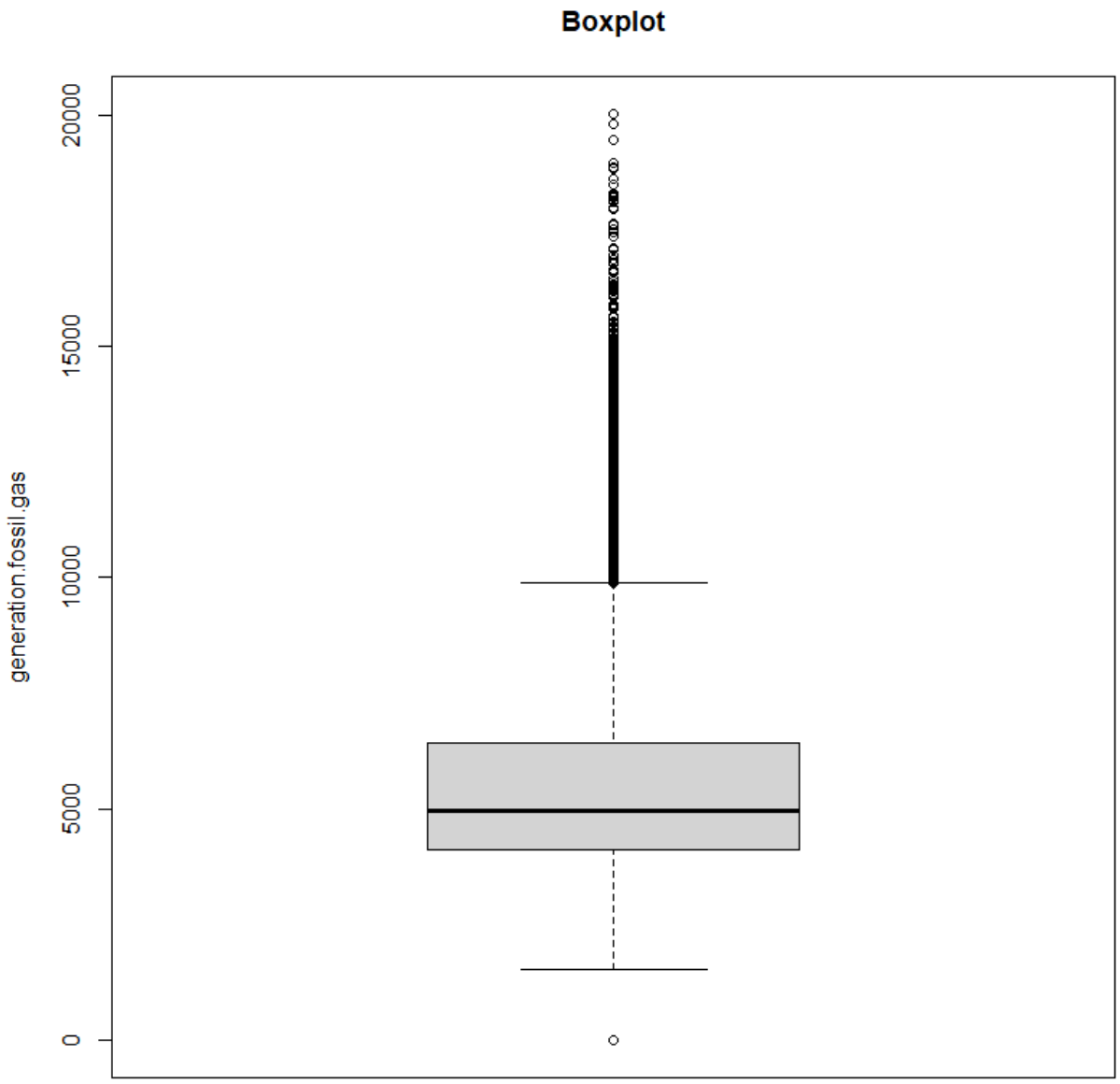
Esios retreived from https://www.esios.ree.es/en/market-and-prices?date=19-05-2023

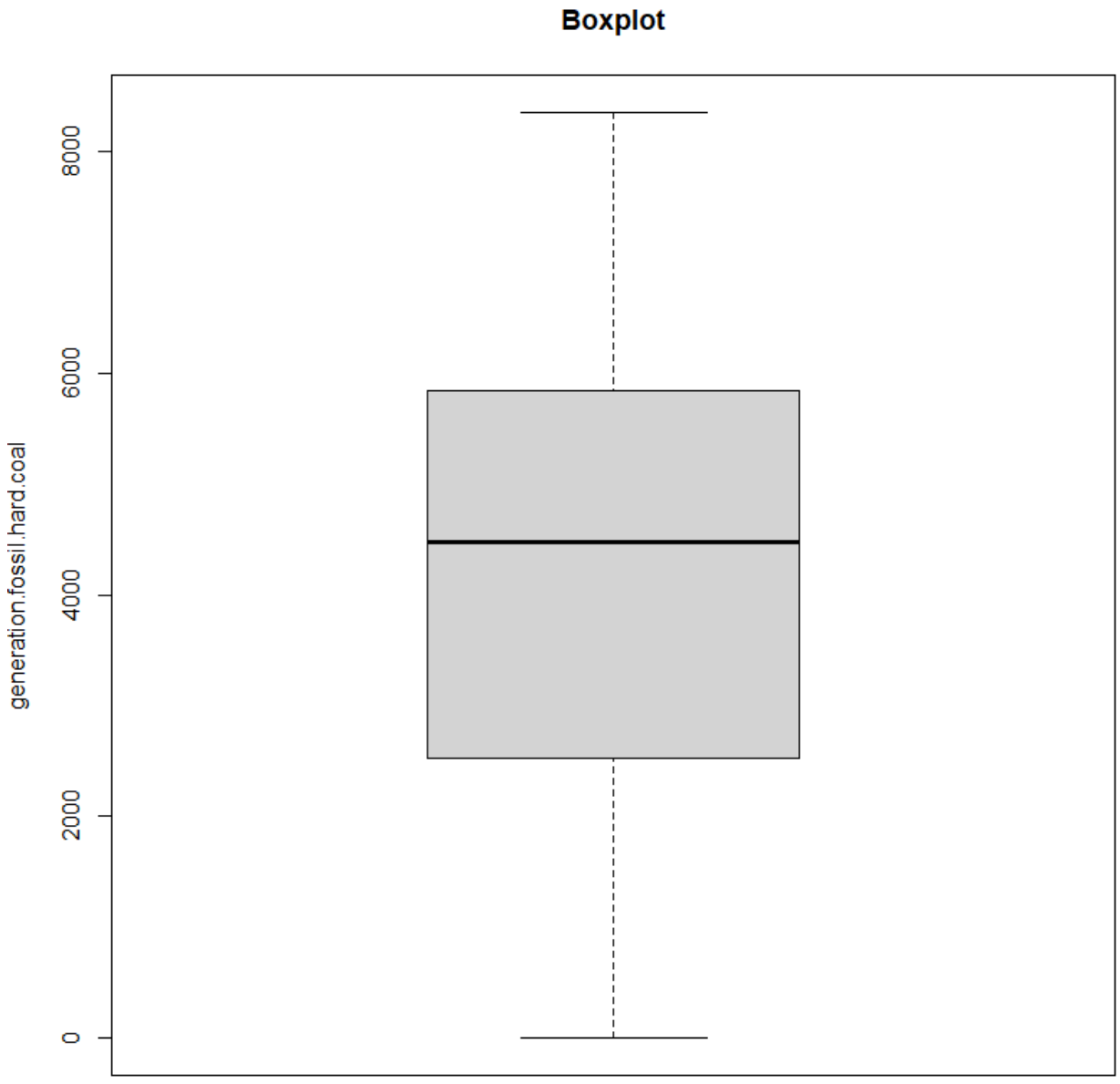
Openweather retrieved from https://openweathermap.org/api

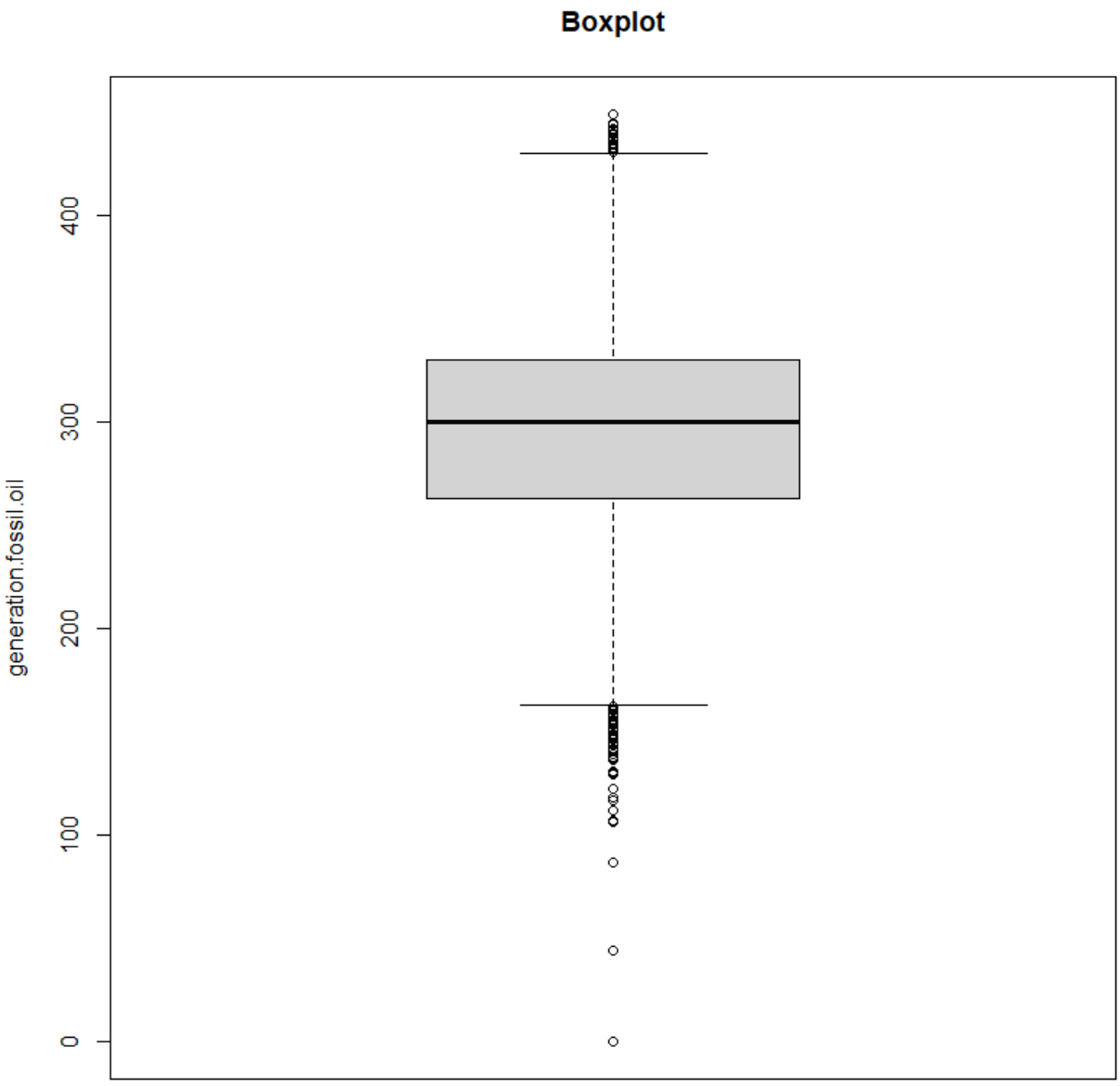
Appendix

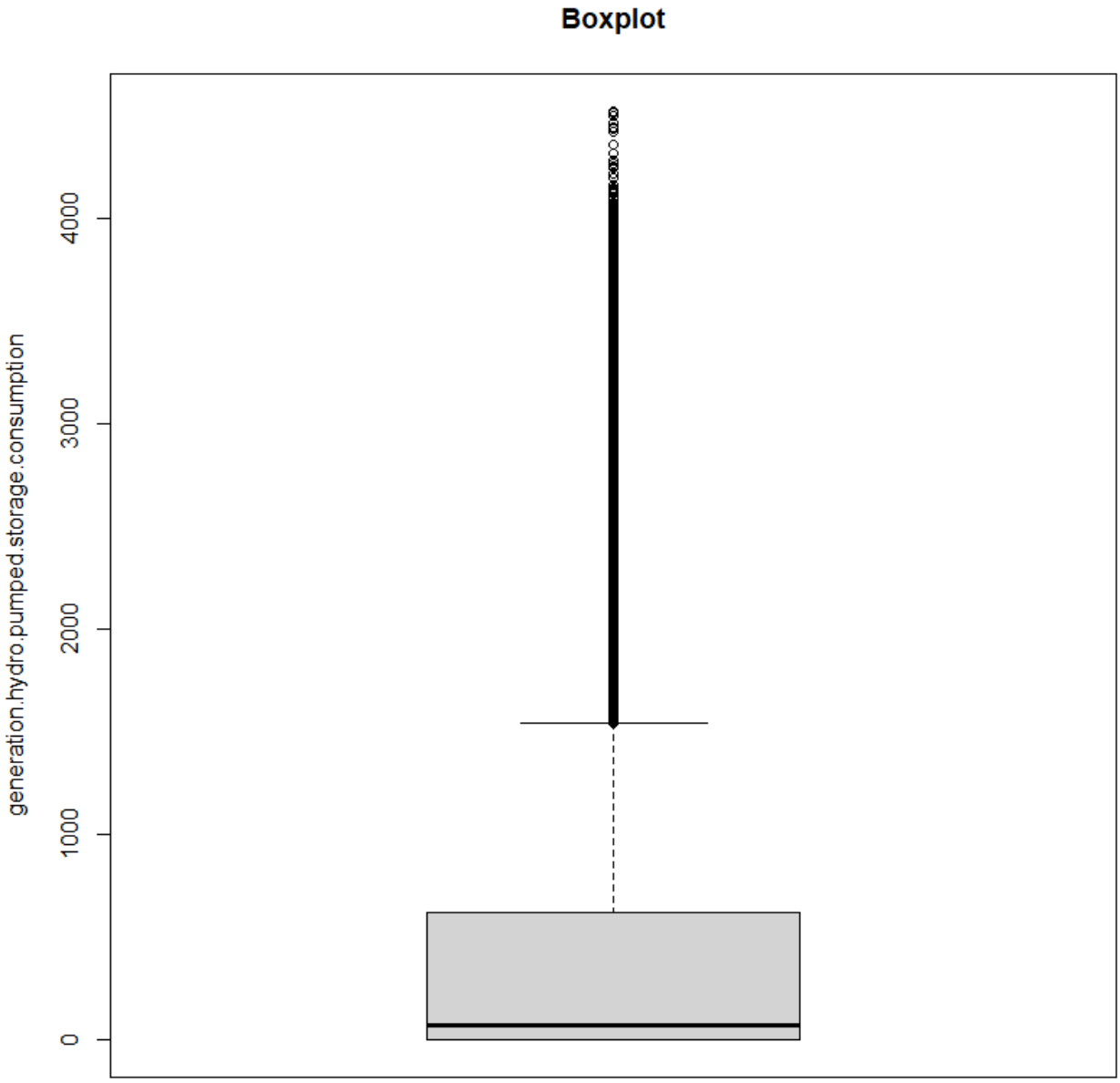


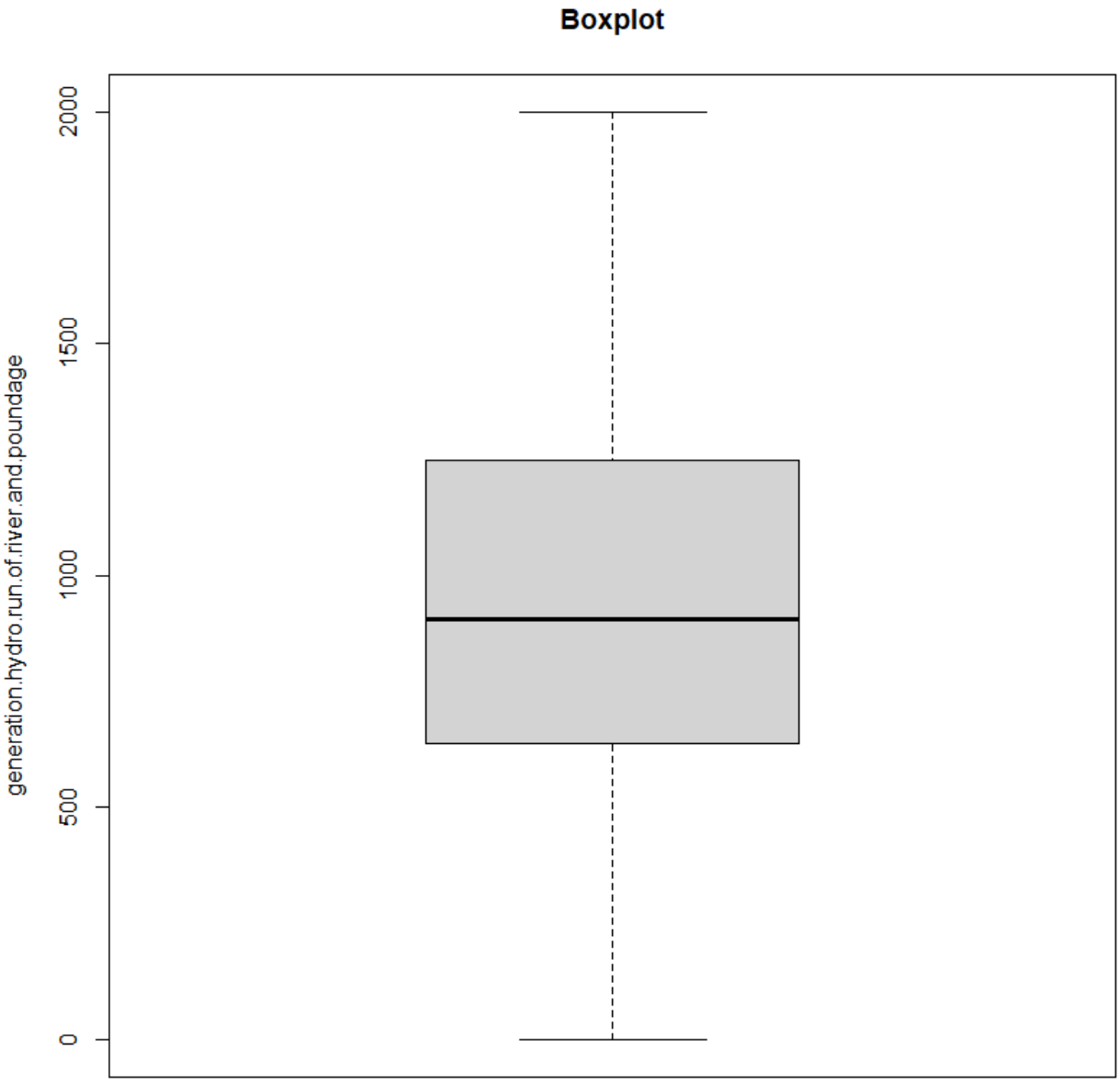

Boxplot
generation.fossil.brown.coal.lignite
0
200
400
600
800
1000

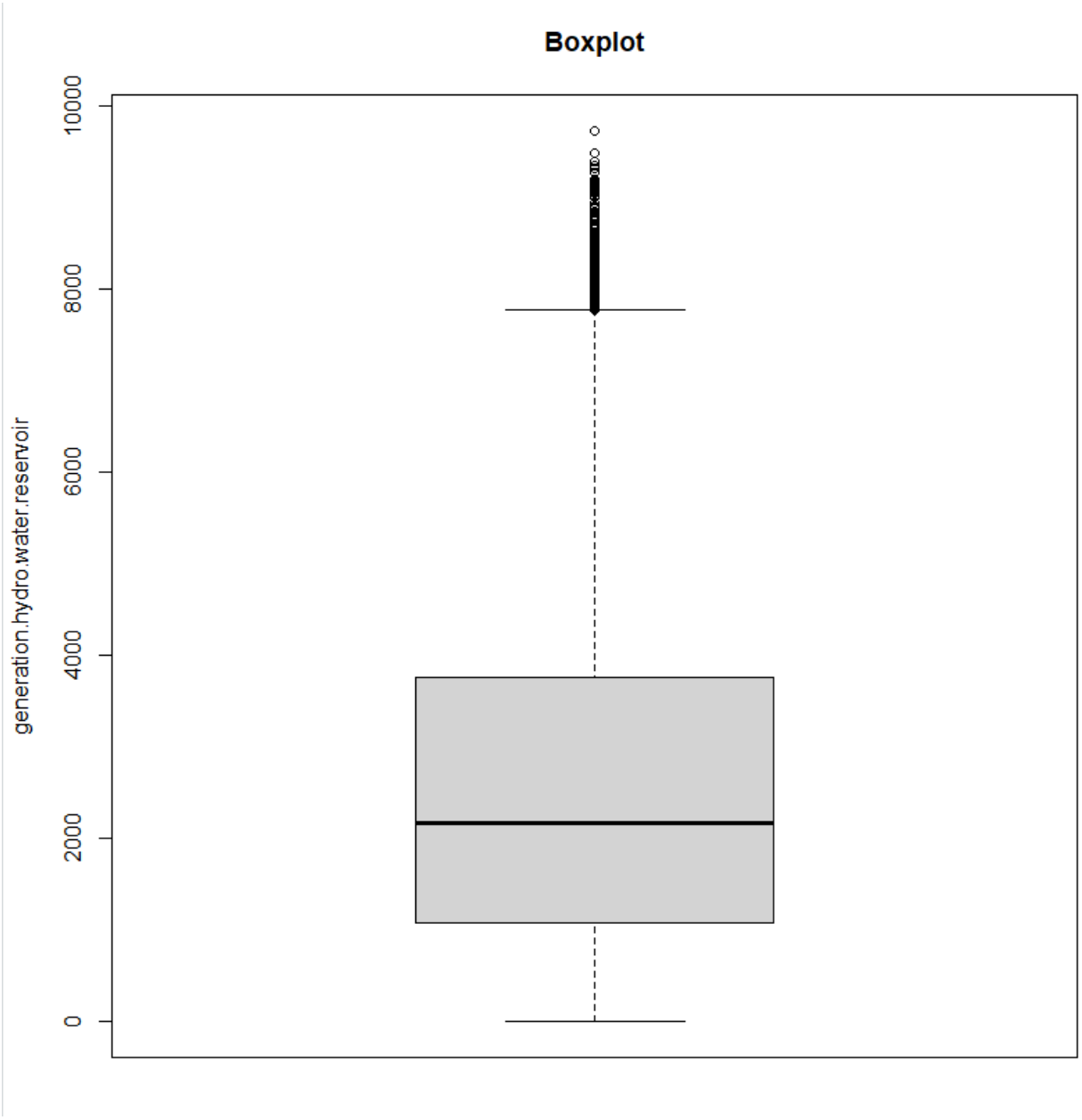

Boxplot
generation.fossil.gas
0
5000
10000
15000
20000

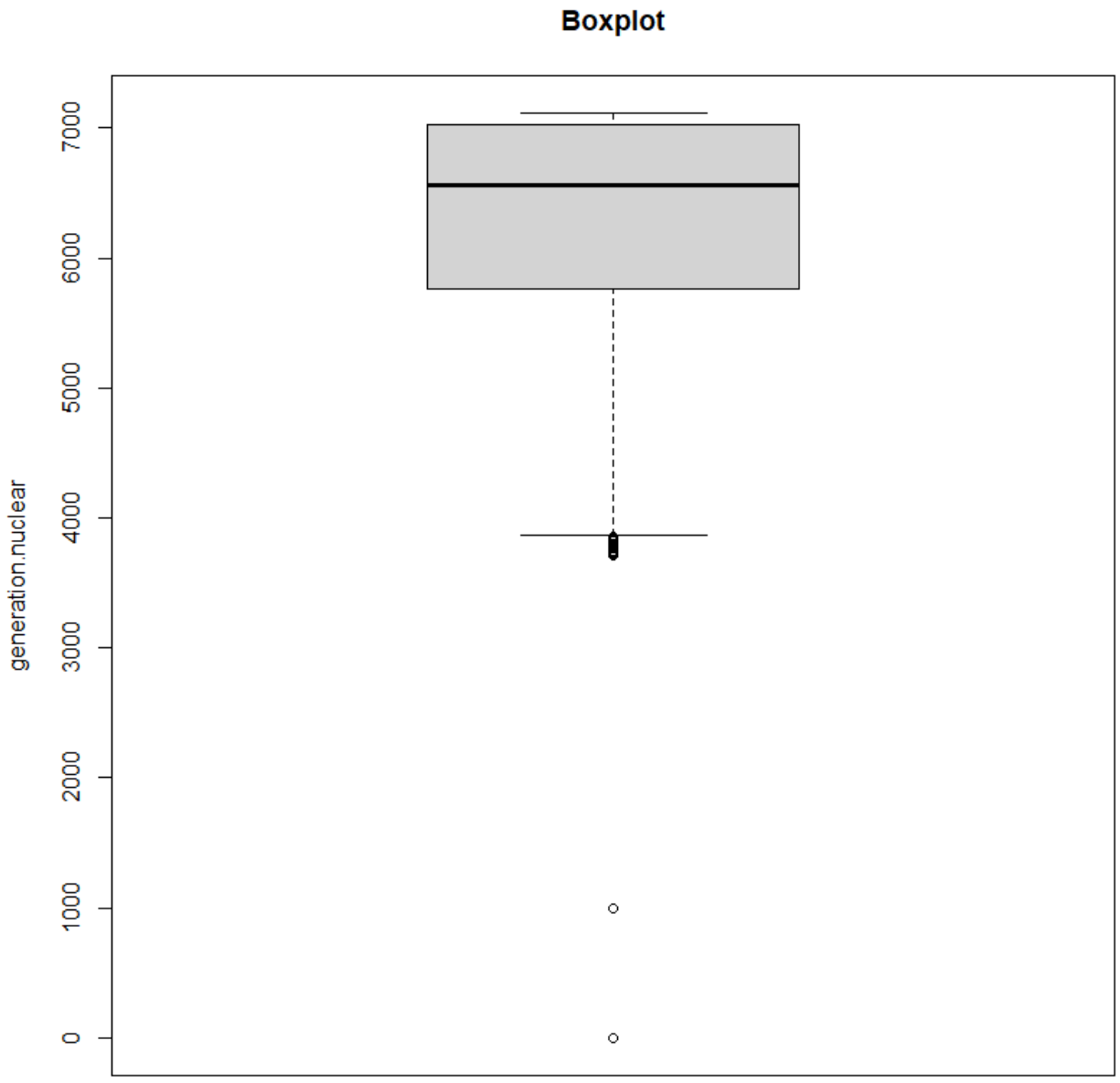

Boxplot
generation.fossil.hard.coal
0
2000
4000
6000
8000

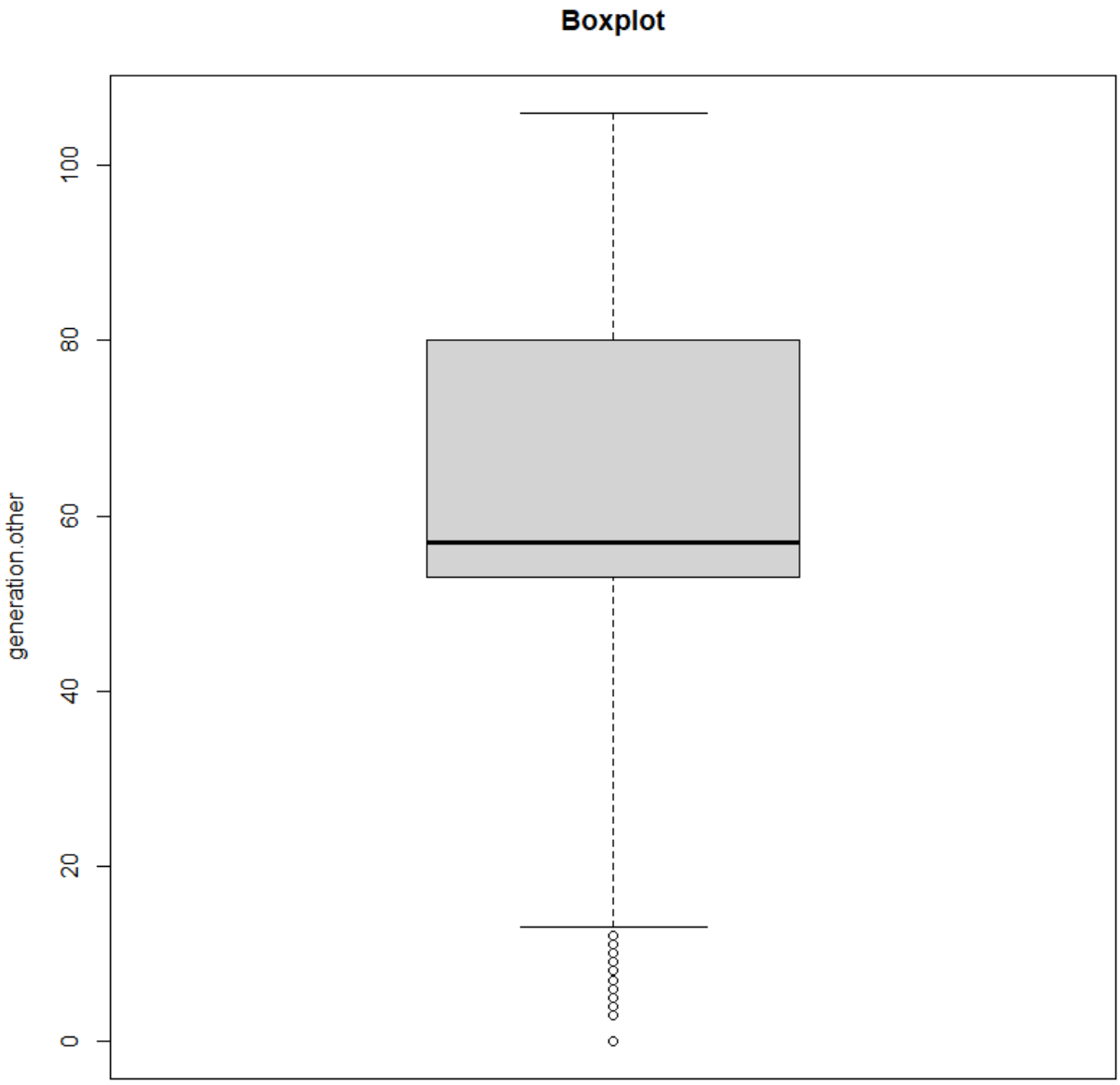

Boxplot
generation.fossil.oil
0
100
200
300
400

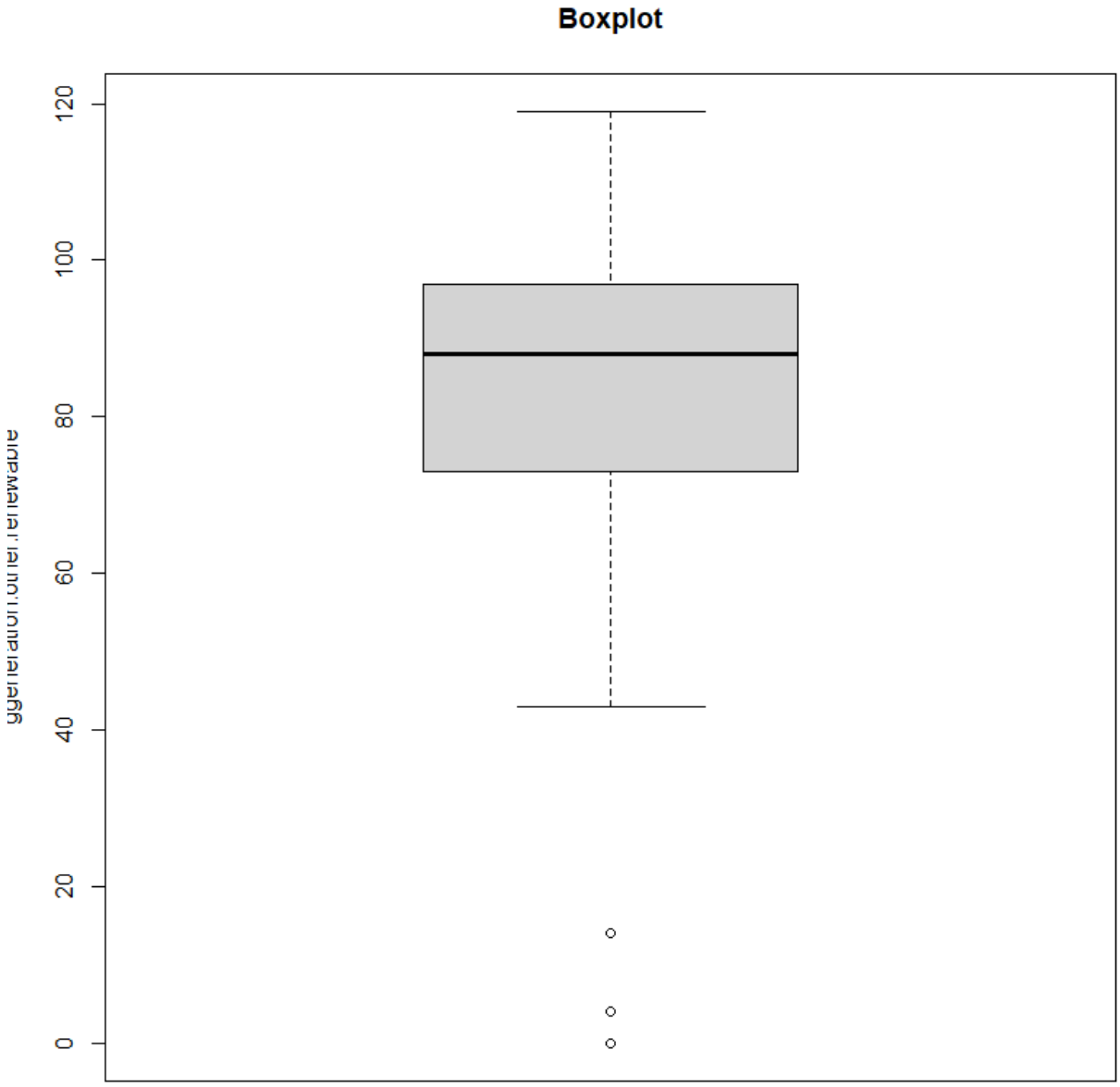

Boxplot
generation.hydro.pumped.storage.consumption
0
1000
2000
3000
4000

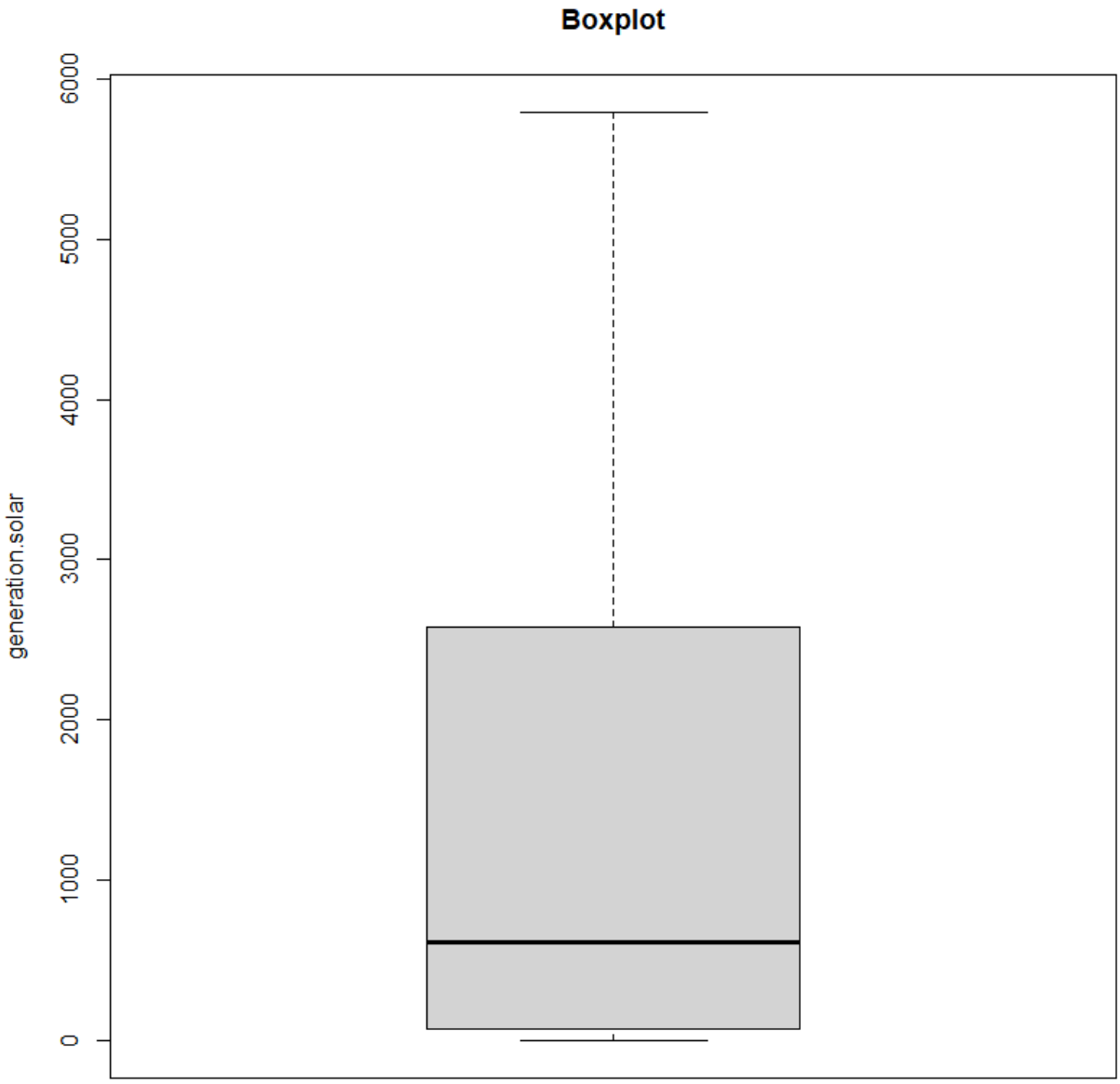

Boxplot
generation.hydro.run.of.river.and.poundage
0
500
1000
1500
2000

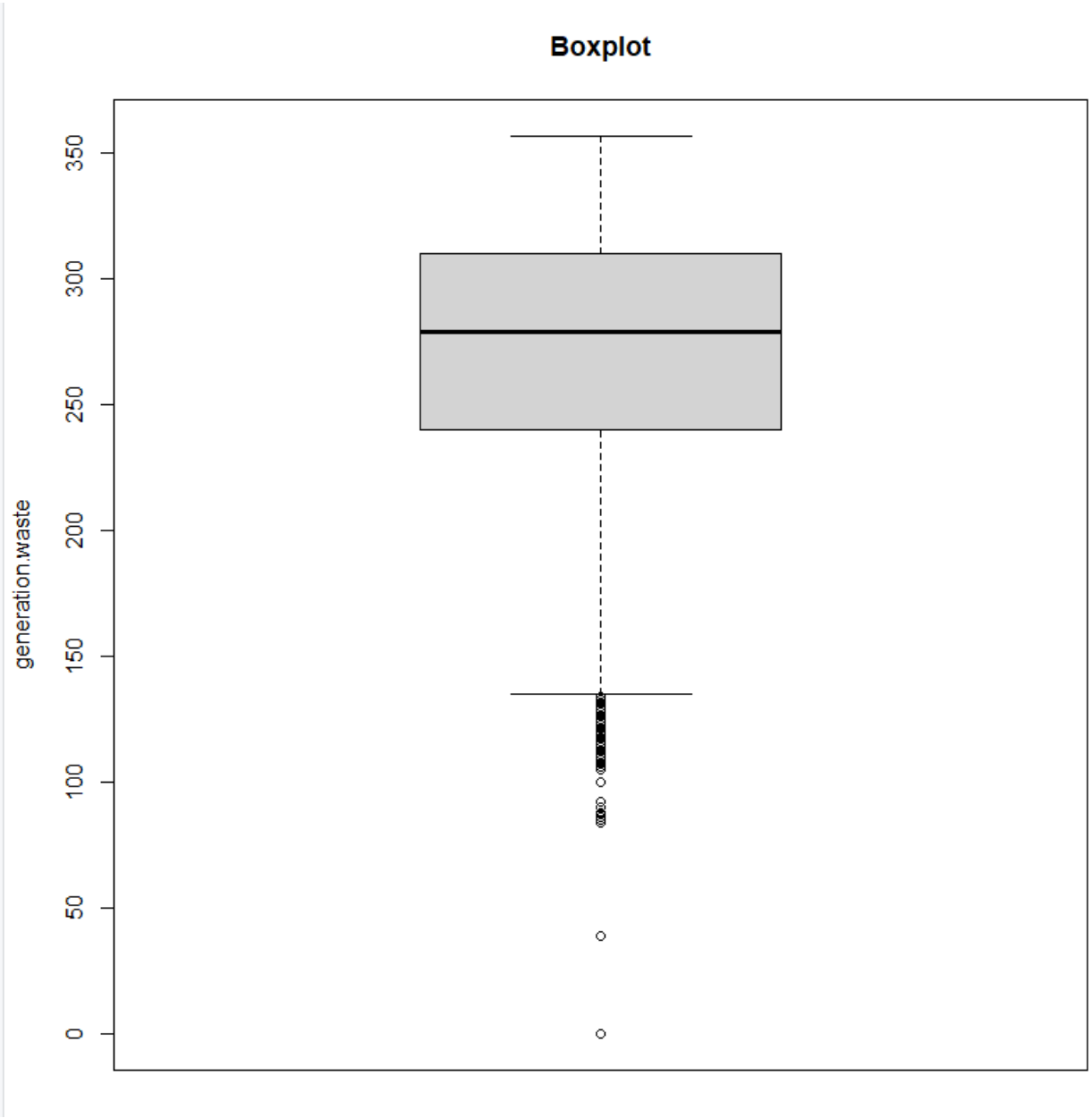

Boxplot
generation.hydro.water.reservoir
0
2000
4000
6000
8000
10000

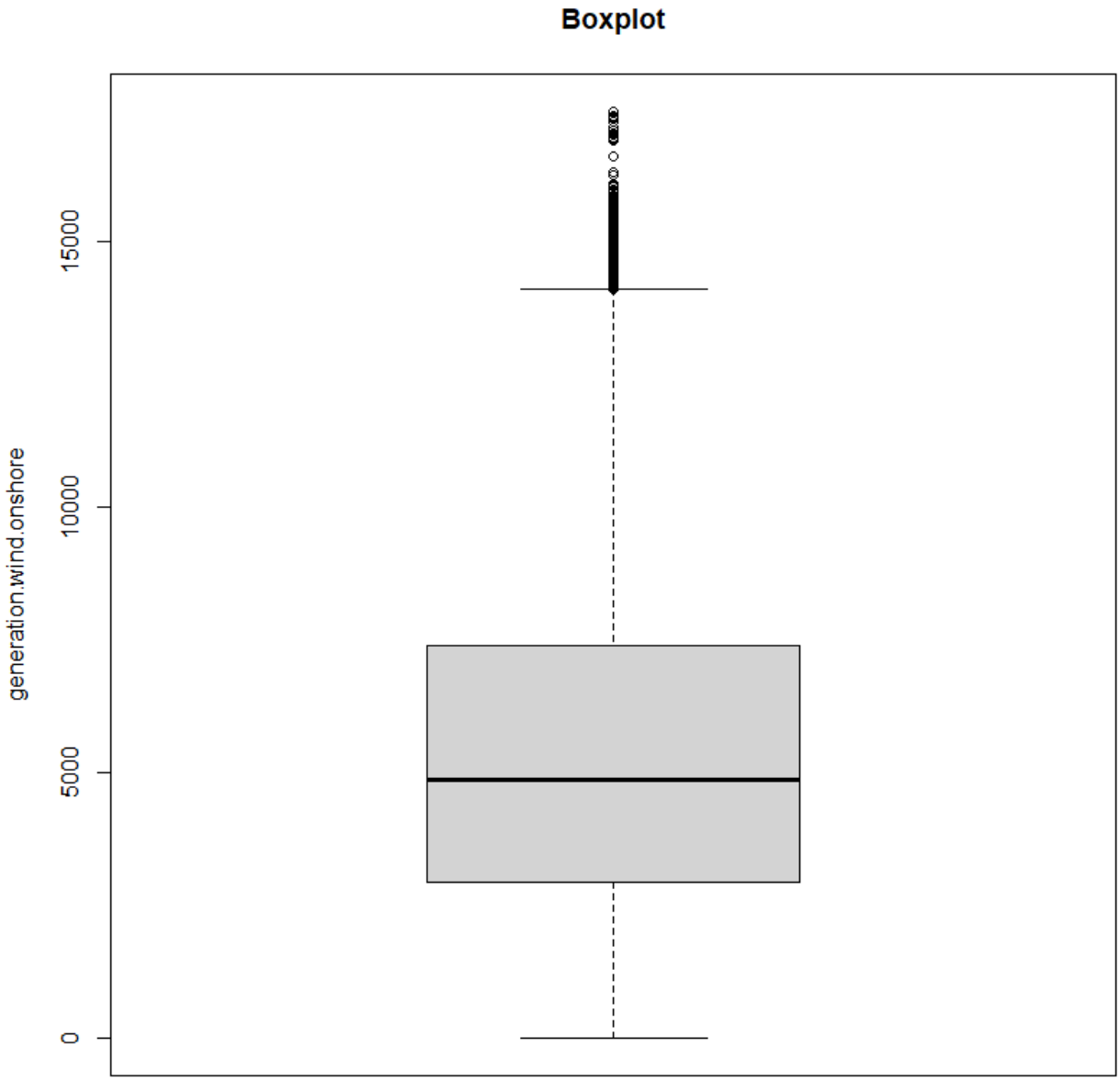

Boxplot
generation.nuclear
0
1000
2000
3000
4000
5000
6000
7000

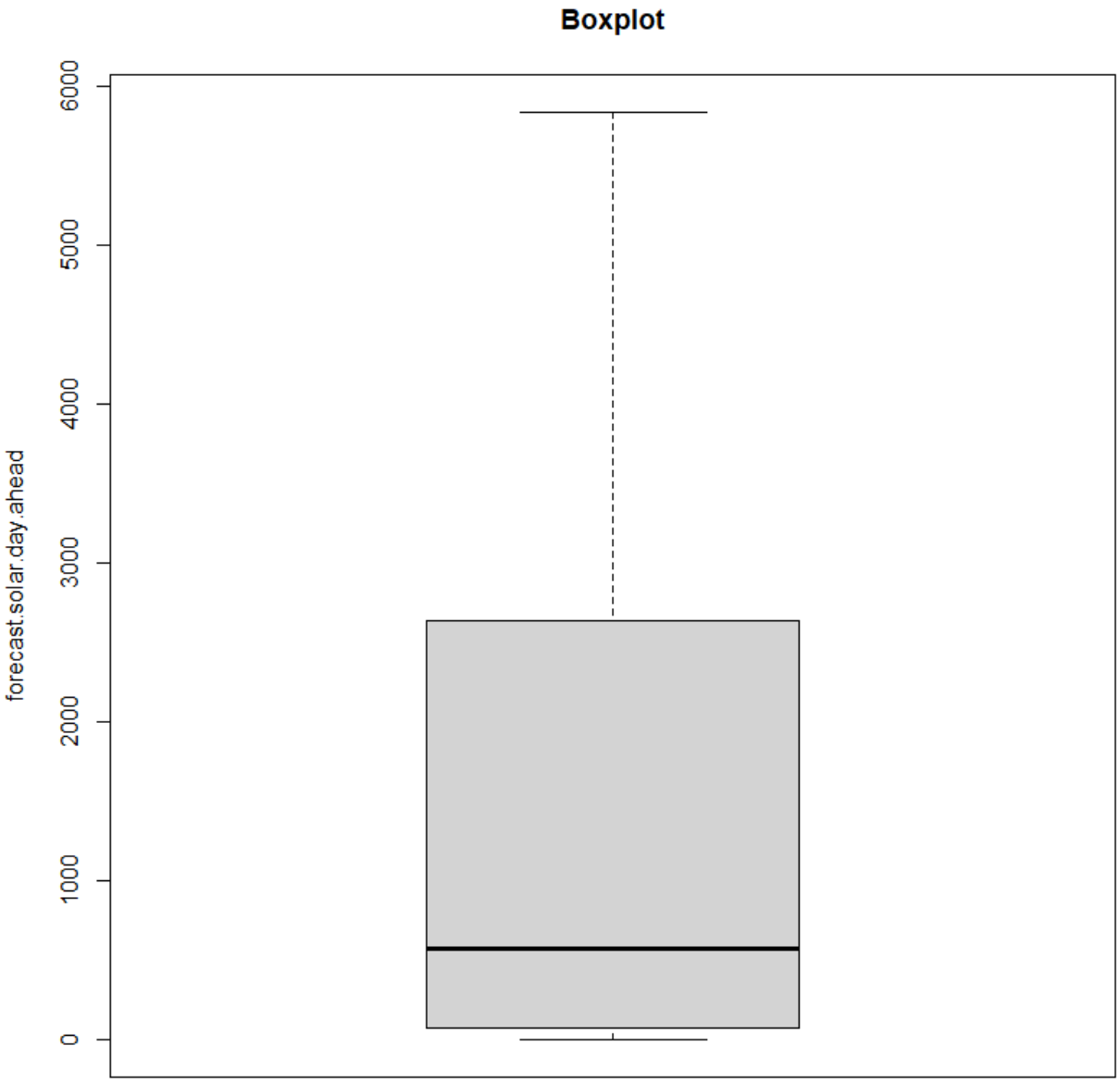

Boxplot
generation.other
0
20
40
60
80
100

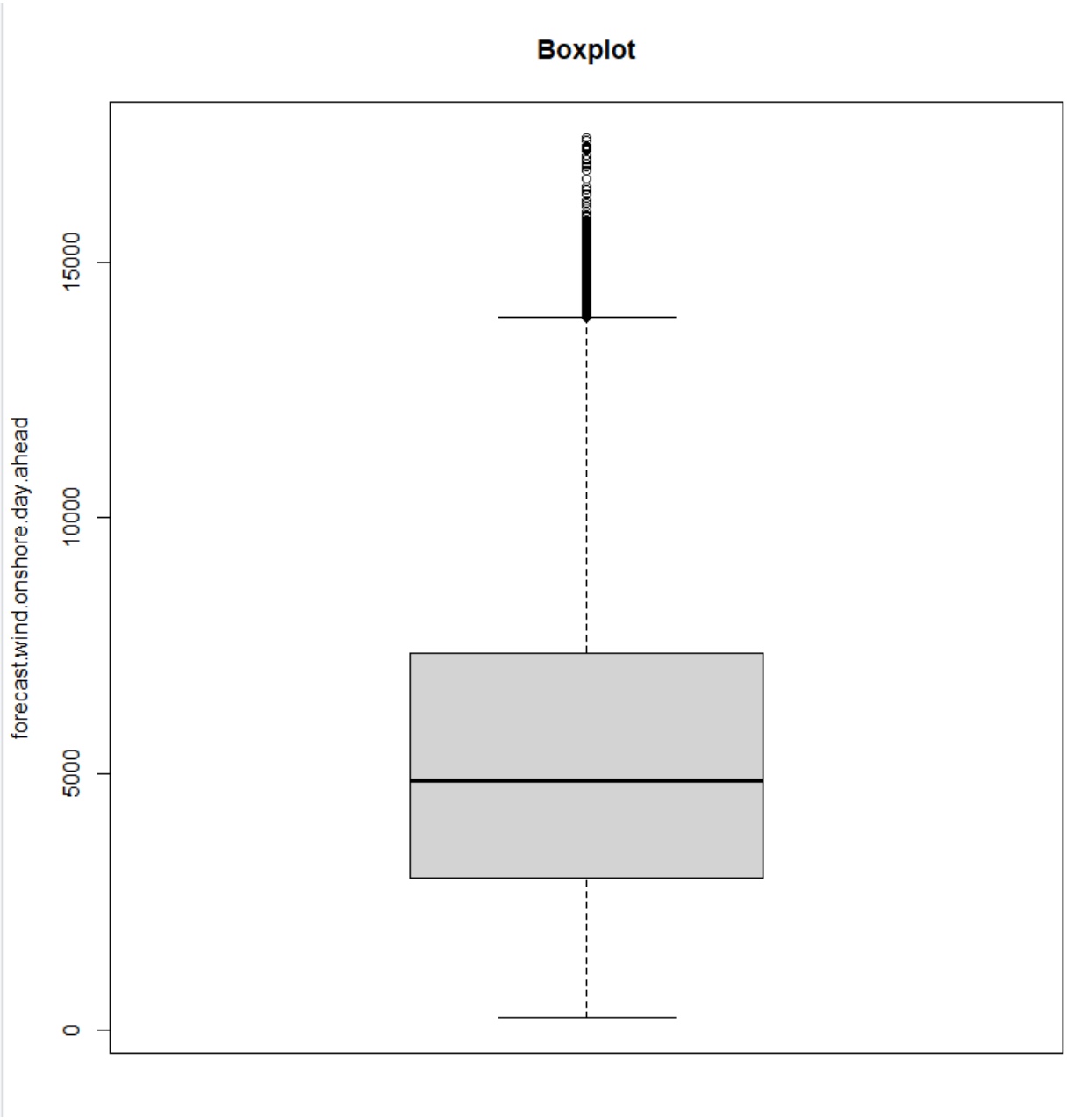

Boxplot
0
20
40
60
80
100
120

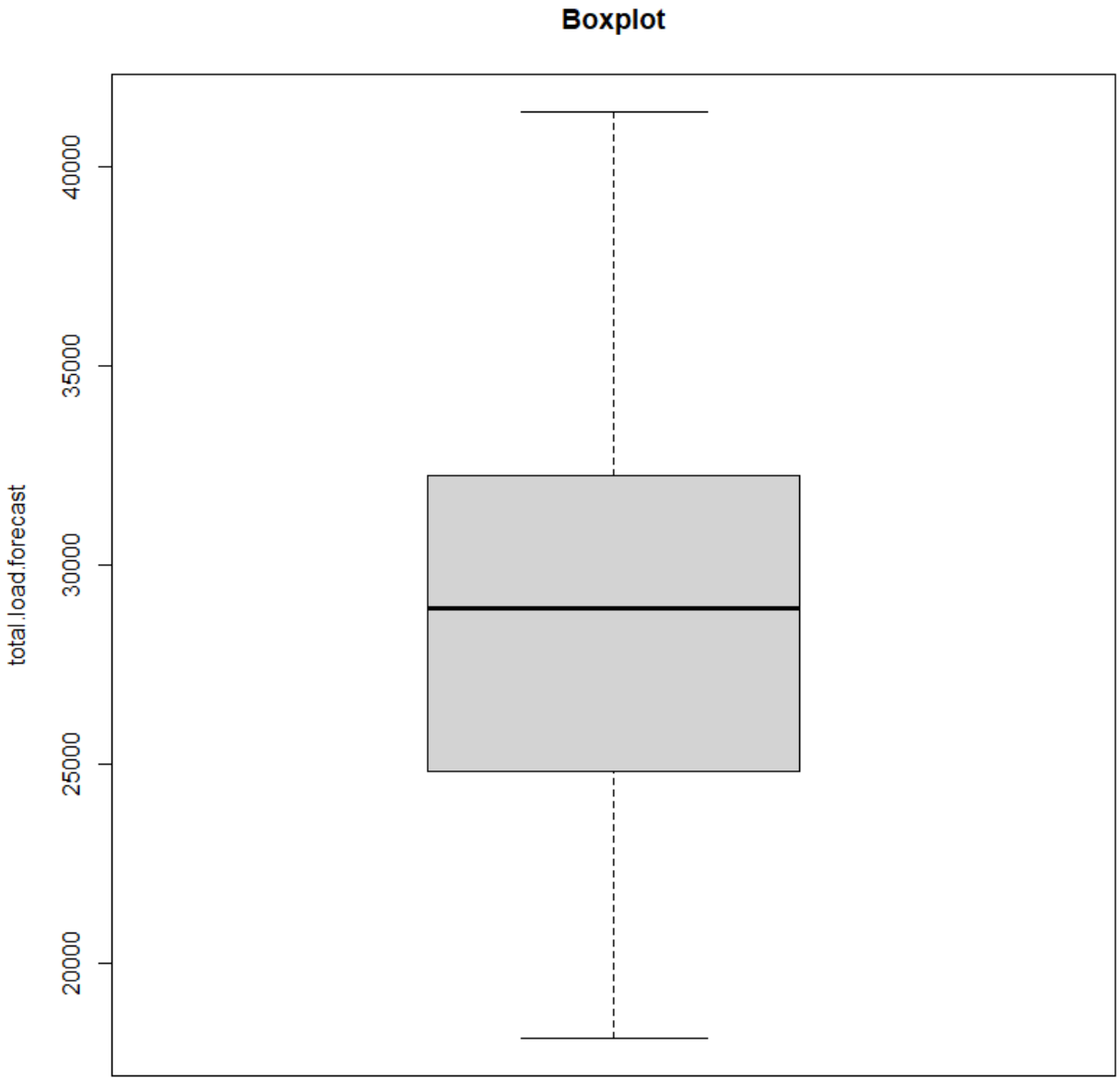

Boxplot
generation.solar
0
1000
2000
3000
4000
5000
6000

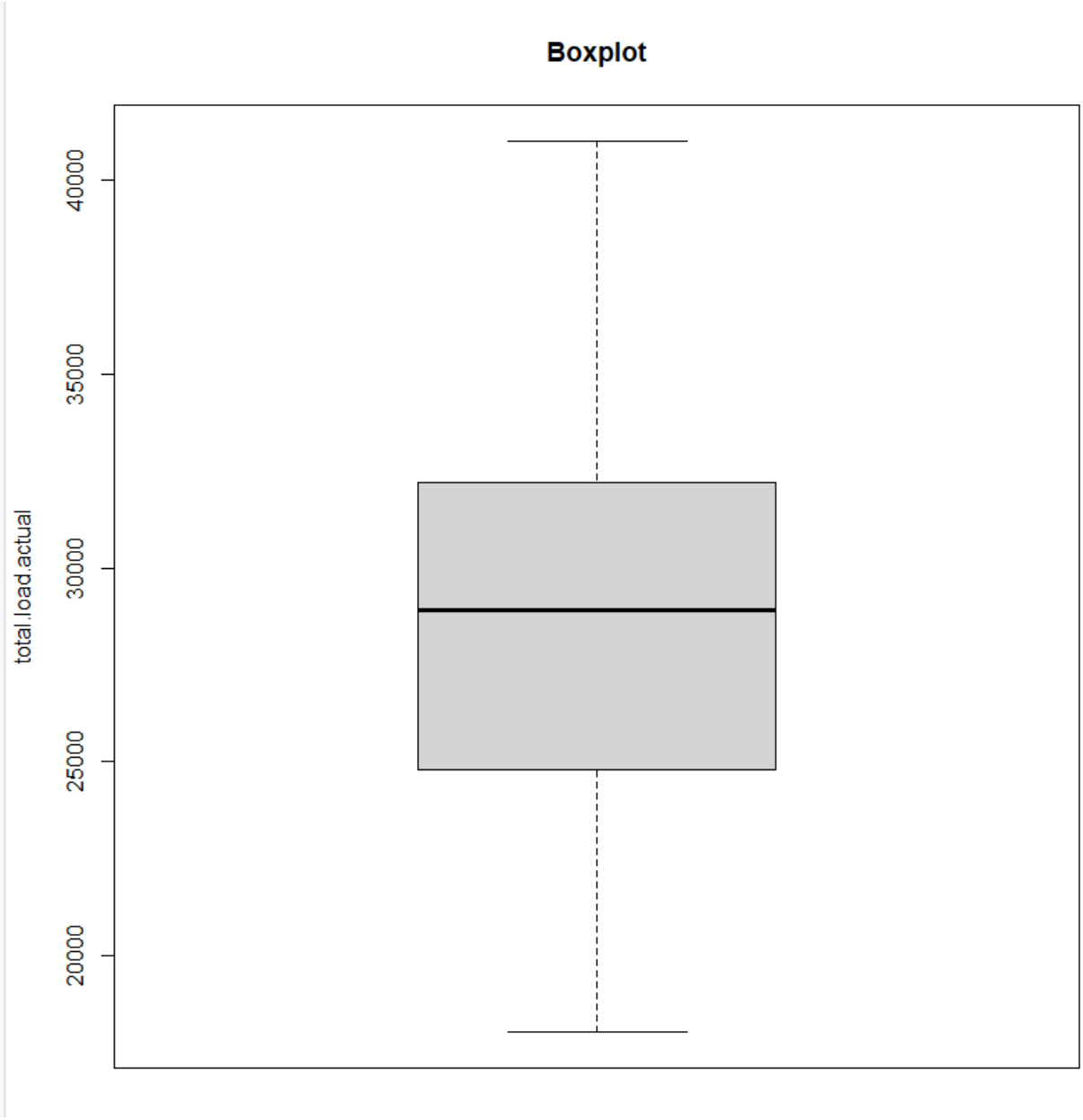

Boxplot
generation.waste
0
50
100
150
200
250
300
350

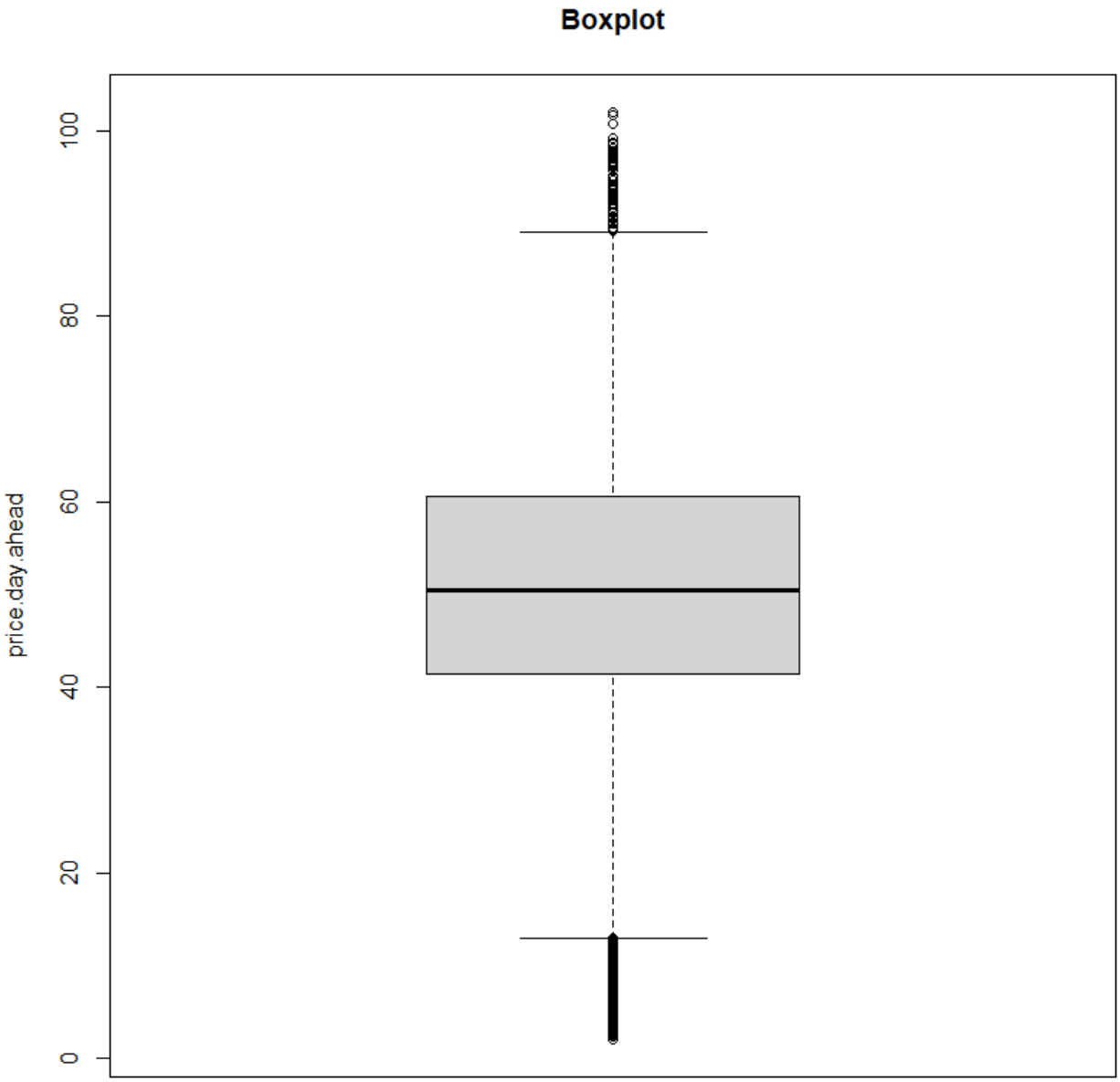

Boxplot
generation.wind.onshore
0
5000
10000
15000

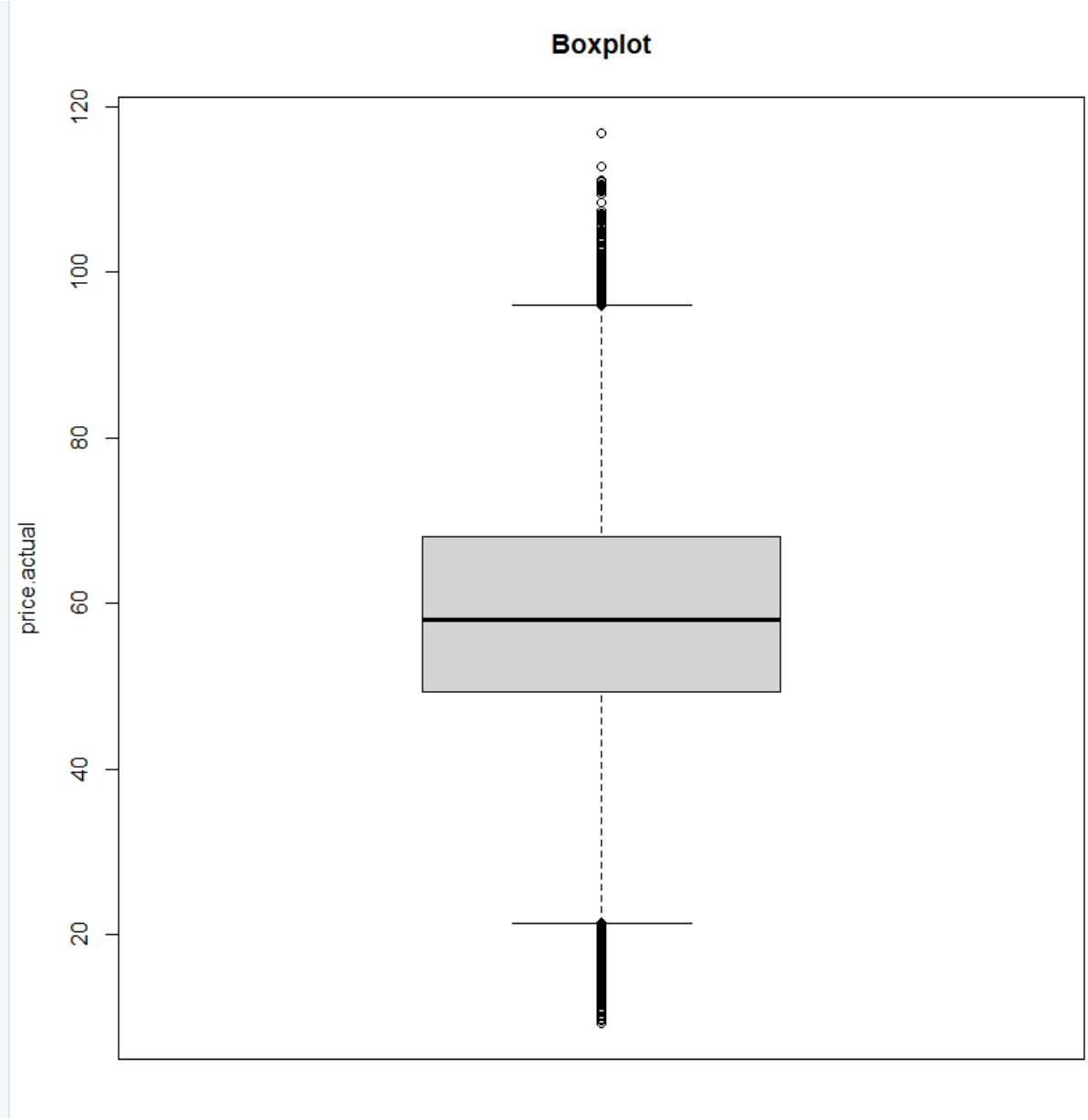

Boxplot
forecast.solar.day.ahead
0
1000
2000
3000
4000
5000
6000

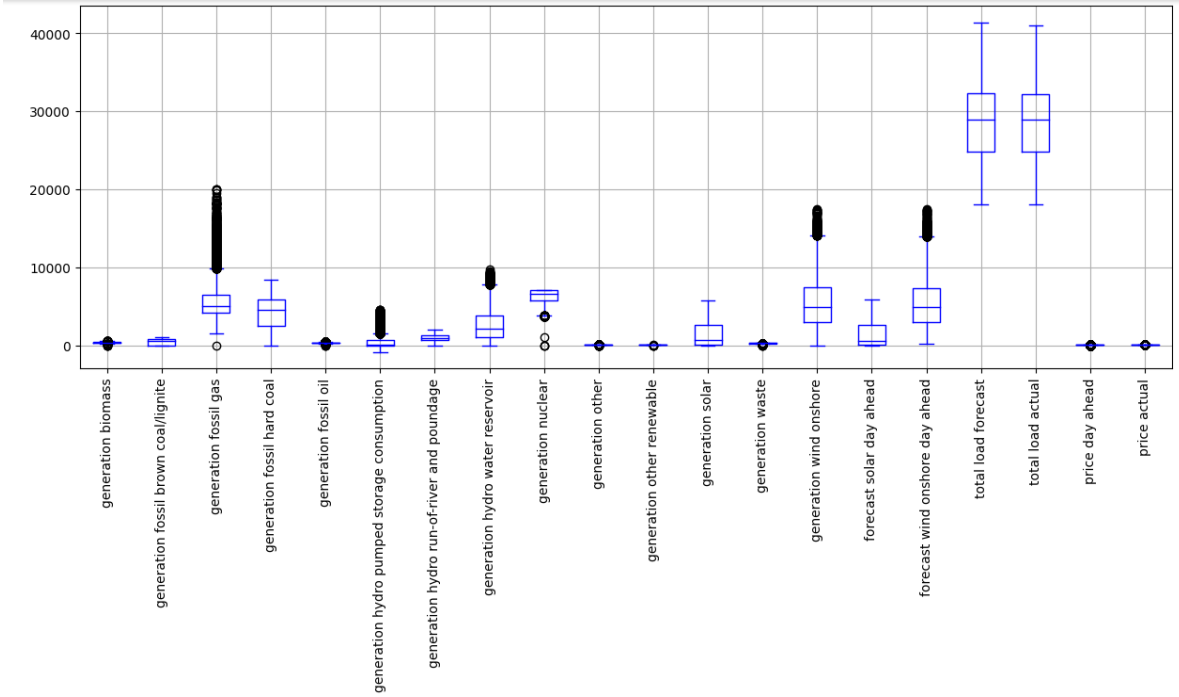

Boxplot
forecast.wind.onshore.day.ahead
0
5000
10000
15000

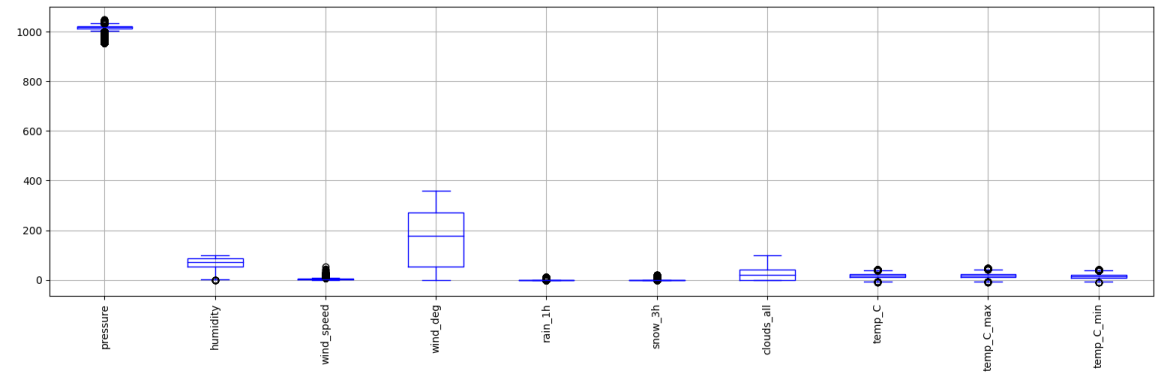

Boxplot
total.load.forecast
20000
25000
30000
35000
40000

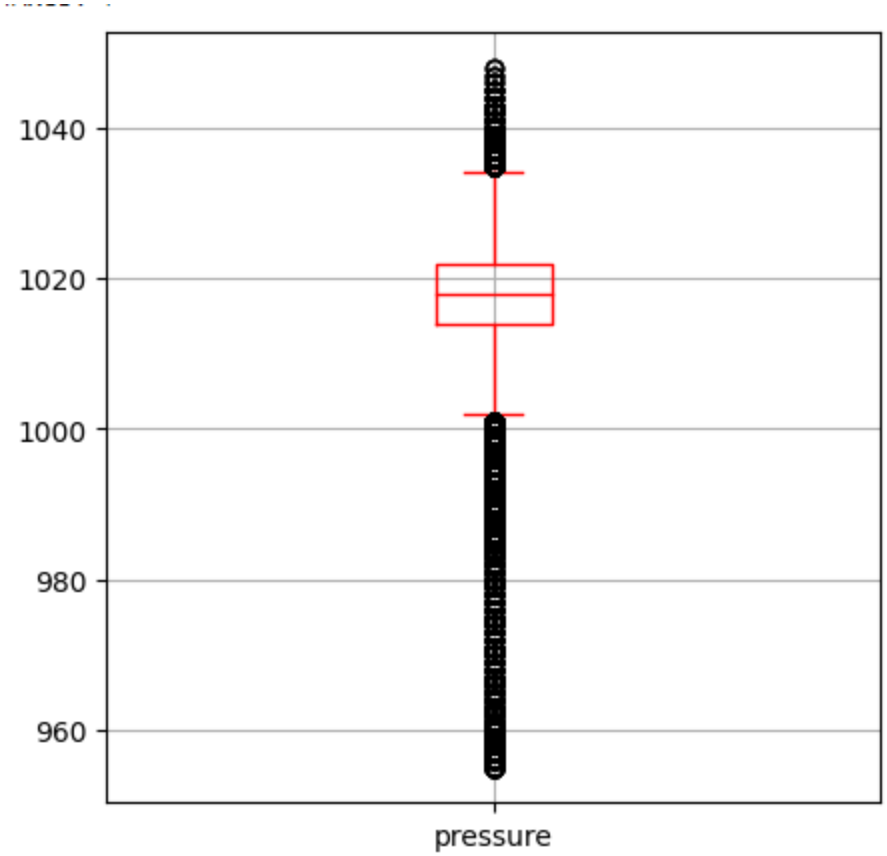

Boxplot
total.load.actual
20000
25000
30000
35000
40000

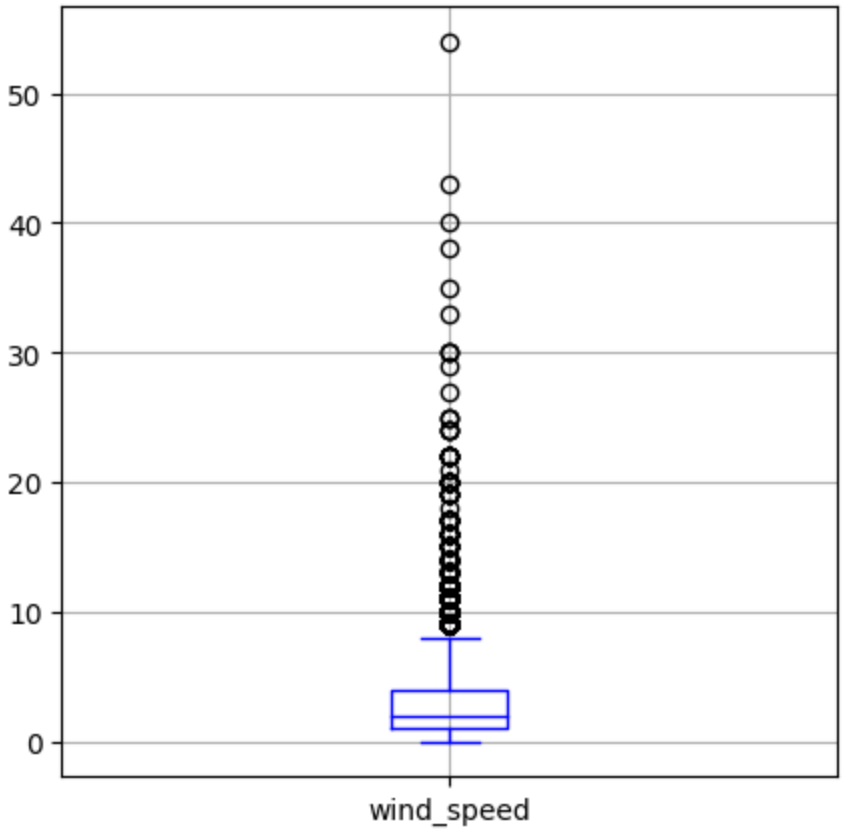

Boxplot
price.day.ahead
0
20
40
60
80
100

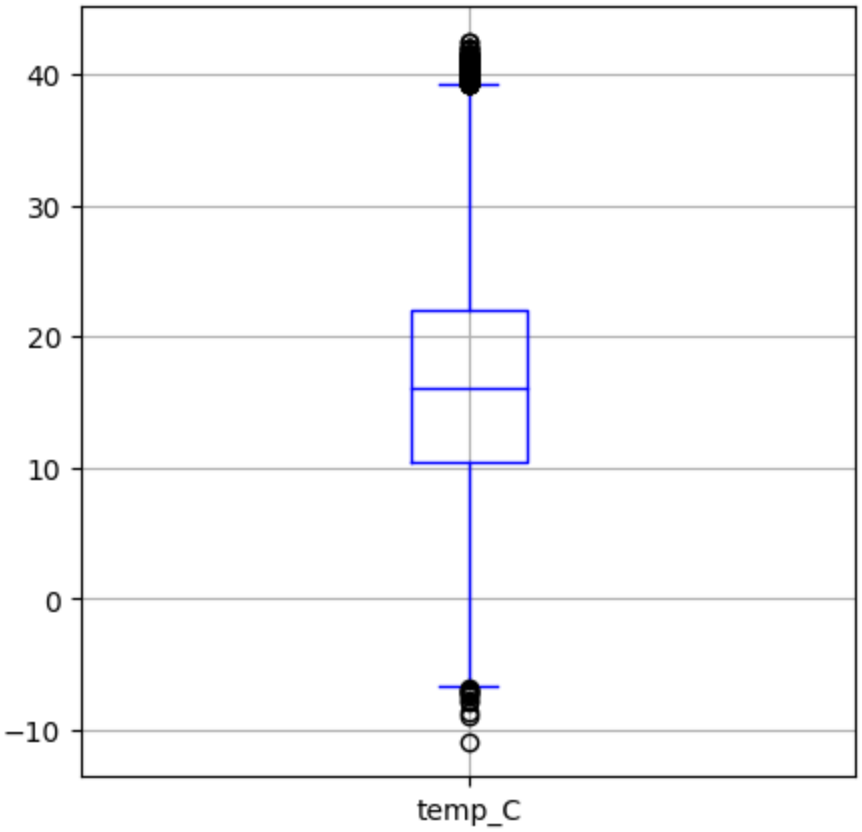

Boxplot
price.actual
120
100
80
60
40
20

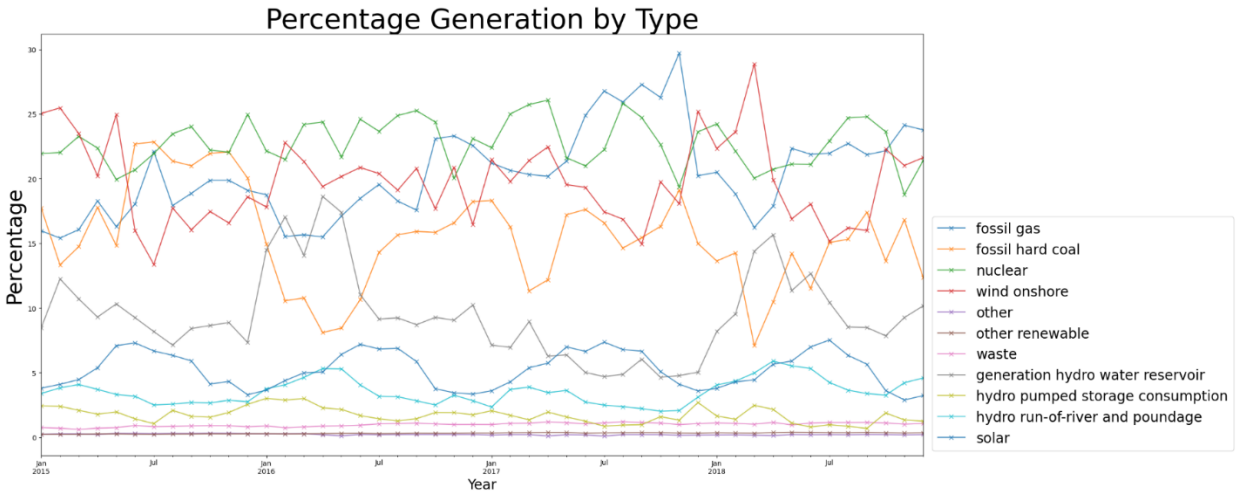

40000
30000
20000
10000
0
generation biomass
generation fossil brown coal/lignite
generation fossil gas
generation fossil hard coal
generation fossil oil
generation hydro pumped storage consumption
generation hydro run-of-river and poundage
generation hydro water reservoir
generation nuclear
generation other
generation other renewable
generation solar
generation waste
generation wind onshore
forecast solar day ahead
forecast wind onshore day ahead
total load forecast
total load actual
price day ahead
price actual

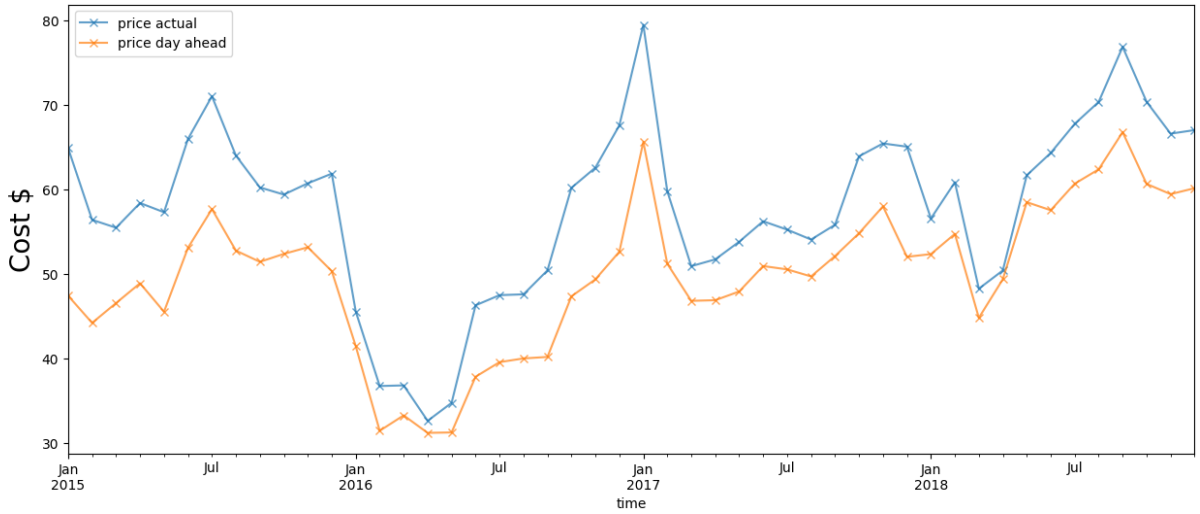

1000
800
600
400
200
0
pressure
humidity
wind_speed
wind_deg
rain_1h
snow_3h
clouds_all
temp_C
temp_C_max
temp_C_min


1040
1020
1000
980
960
pressure

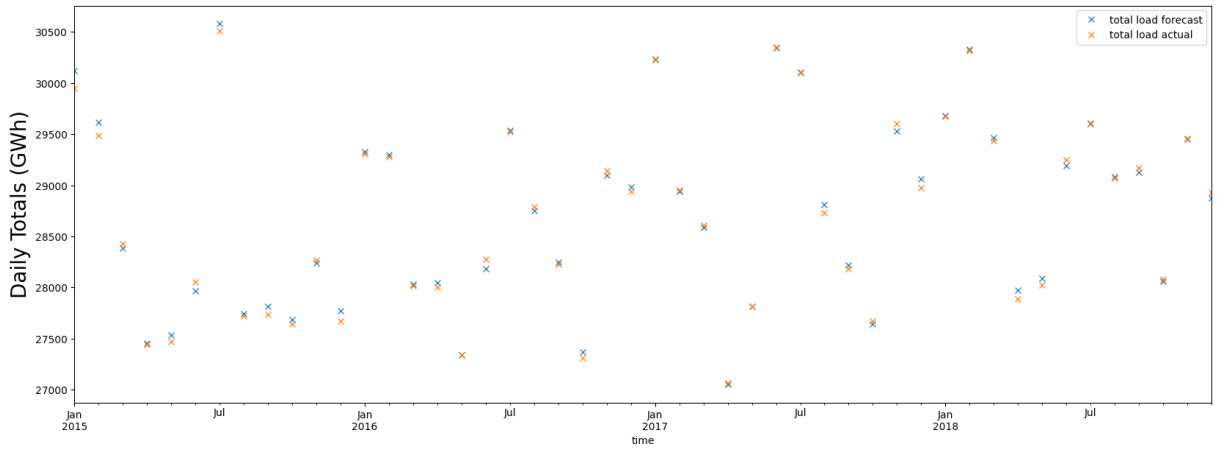

50
40
30
20
10
0
wind_speed

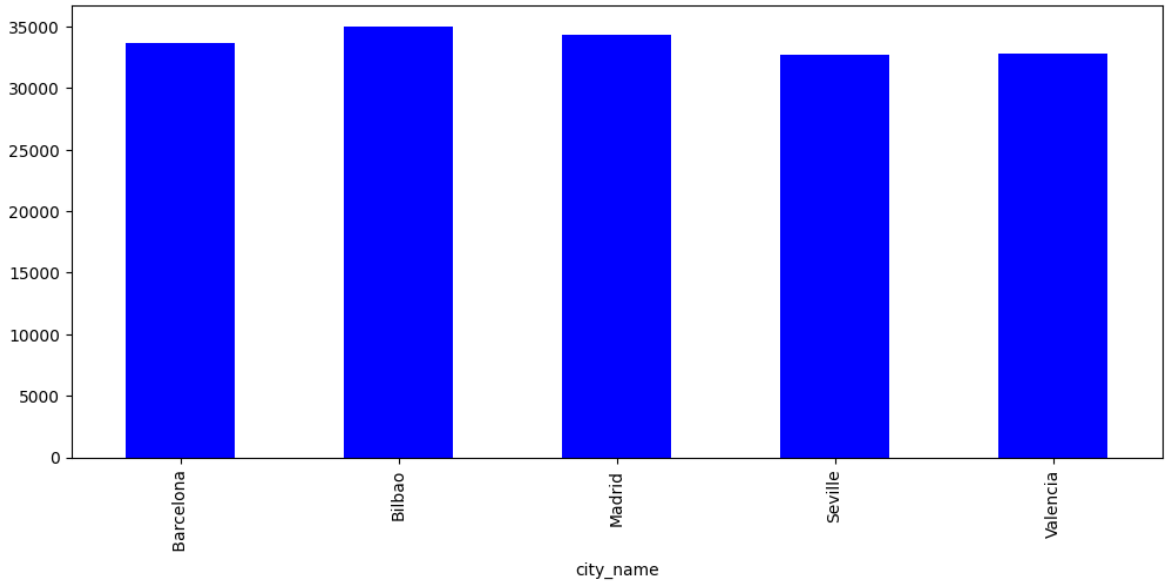

40
30
20
10
0
−10
temp_C

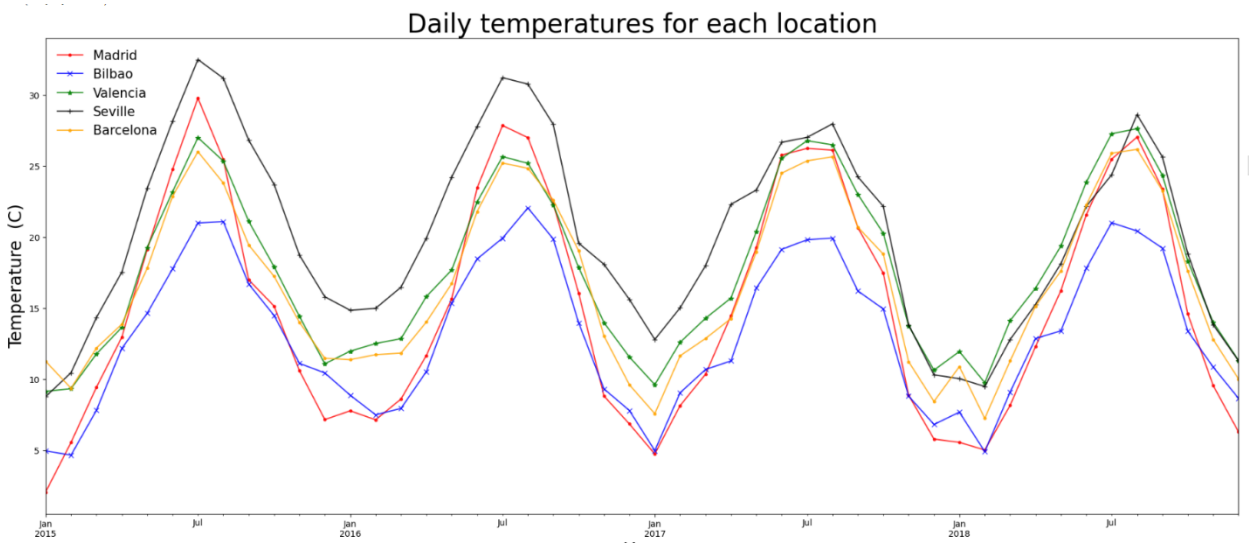

Percentage Generation by Type
Percentage
Year
fossil gas
fossil hard coal
nuclear
wind onshore
other
other renewable
waste
generation hydro water reservoir
hydro pumped storage consumption
hydro run-of-river and poundage
solar

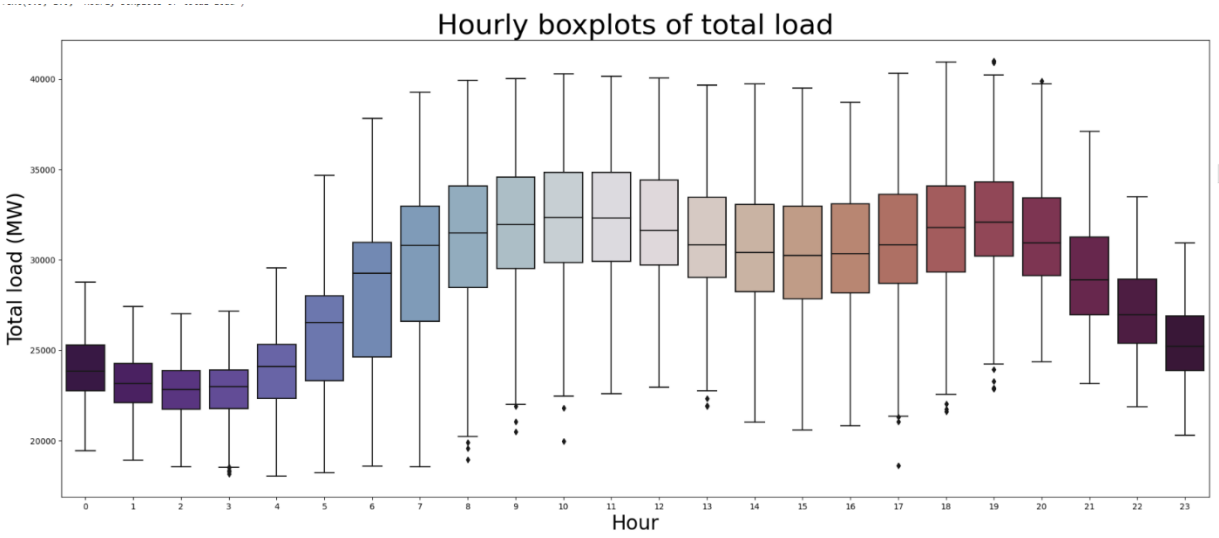

price actual
price day ahead
Cost $
time
Jan 2015
Jul
Jan 2016
Jul
Jan 2017
Jul
Jan 2018
Jul

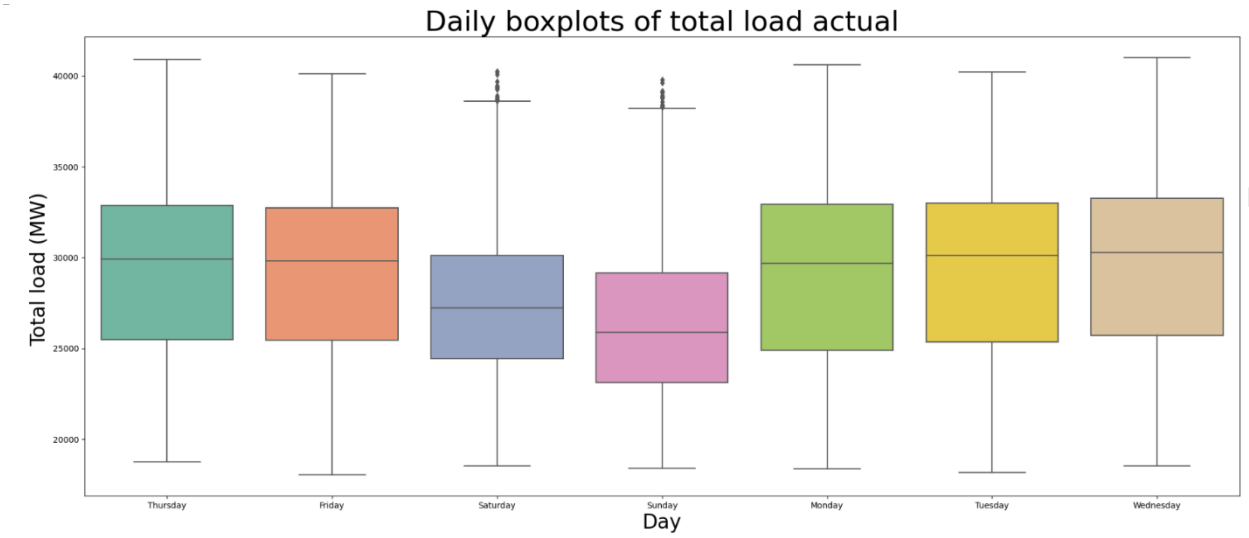
Predicting Energy Demand in Spain

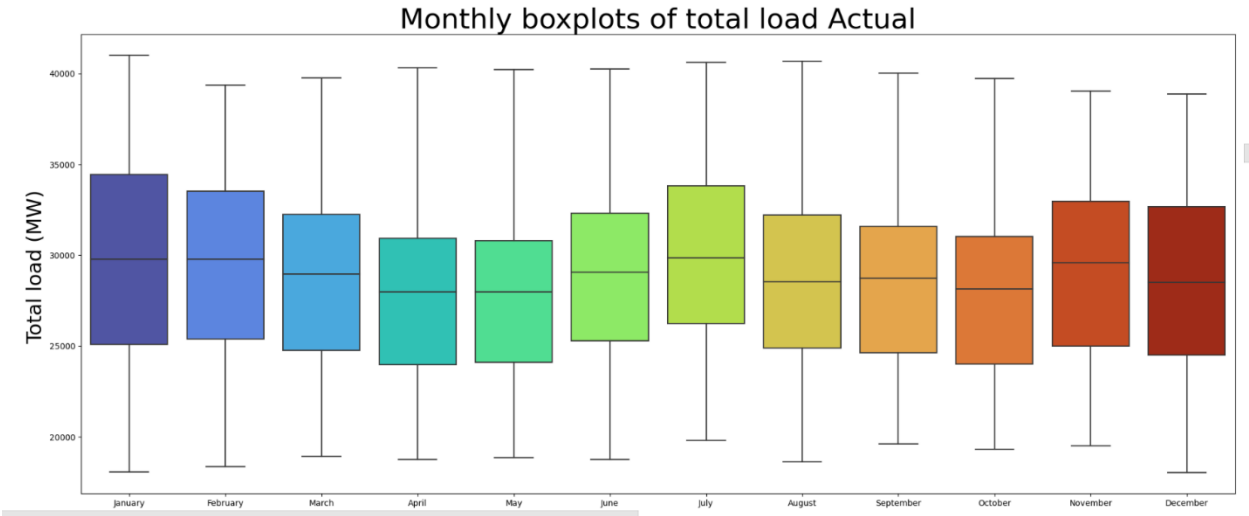





Daily temperatures for each location
Madrid
Bilbao
Valencia
Seville
Barcelona
Temperature (C)
Jan 2015
Jul
Jan 2016
Jul
Jan 2017
Jul
Jan 2018
Jul


Hourly boxplots of total load
Total load (MW)
Hour

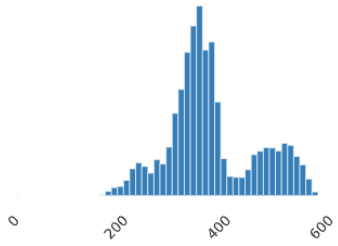

Daily boxplots of total load actual
Total load (MW)
40000
35000
30000
25000
20000
Thursday
Friday
Saturday
Sunday
Monday
Tuesday
Wednesday
Day


Monthly boxplots of total load Actual
Total load (MW)
40000
35000
30000
25000
20000
January
February
March
April
May
June
July
August
September
October
November
December

generation biomass

Real number (ℝ)

Distinct	423	**Minimum**	0
Distinct (%)	1.2%	**Maximum**	592
Missing	19	**Zeros**	4
Missing (%)	0.1%	**Zeros (%)**	< 0.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	383.51173	**Memory size**	547.9 KiB

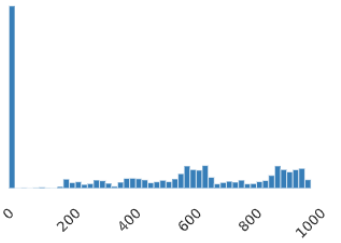


generation fossil brown coal/lignite

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	956	**Minimum**	0
Distinct (%)	2.7%	**Maximum**	999
Missing	18	**Zeros**	10517
Missing (%)	0.1%	**Zeros (%)**	30.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	448.06261	**Memory size**	547.9 KiB



generation fossil coal-derived gas

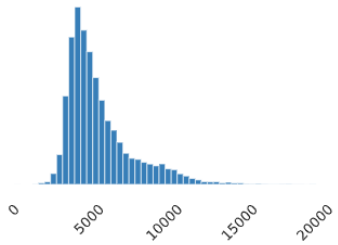
Categorical

Distinct	1
Distinct (%)	< 0.1%
Missing	18
Missing (%)	0.1%
Memory size	547.9 KiB

generation fossil gas

Real number (ℝ)

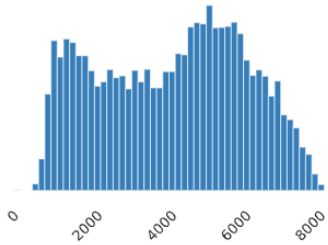
Distinct	8297	**Minimum**	0
Distinct (%)	23.7%	**Maximum**	20034
Missing	18	**Zeros**	1
Missing (%)	0.1%	**Zeros (%)**	< 0.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	5622.7597	**Memory size**	547.9 KiB



generation fossil hard coal

Real number (ℝ)

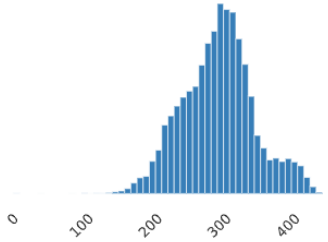
Distinct	7266	**Minimum**	0
Distinct (%)	20.7%	**Maximum**	8359
Missing	18	**Zeros**	3
Missing (%)	0.1%	**Zeros (%)**	< 0.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	4256.0496	**Memory size**	547.9 KiB



generation fossil oil

Real number (ℝ)

Distinct	321	**Minimum**	0
Distinct (%)	0.9%	**Maximum**	449
Missing	19	**Zeros**	3
Missing (%)	0.1%	**Zeros (%)**	< 0.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	298.32368	**Memory size**	547.9 KiB



generation fossil oil shale

Categorical

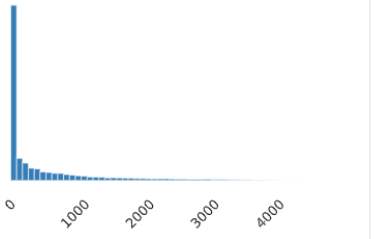
Distinct	1
Distinct (%)	< 0.1%
Missing	18
Missing (%)	0.1%
Memory size	547.9 KiB

generation hydro pumped storage consumption

Real number (ℝ)

HIGH CORRELATION ZEROS

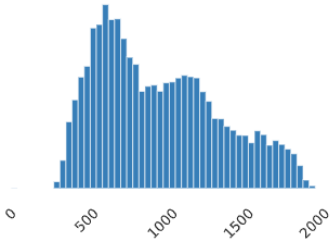
Distinct	3311	**Minimum**	0
Distinct (%)	9.4%	**Maximum**	4523
Missing	19	**Zeros**	12607
Missing (%)	0.1%	**Zeros (%)**	36.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	475.56629	**Memory size**	547.9 KiB



generation hydro run-of-river and poundage

Real number (ℝ)

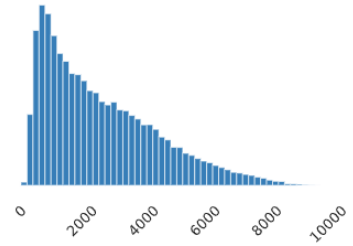
Distinct	1684	**Minimum**	0
Distinct (%)	4.8%	**Maximum**	2000
Missing	19	**Zeros**	3
Missing (%)	0.1%	**Zeros (%)**	< 0.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	972.11386	**Memory size**	547.9 KiB



generation hydro water reservoir

Real number (ℝ)

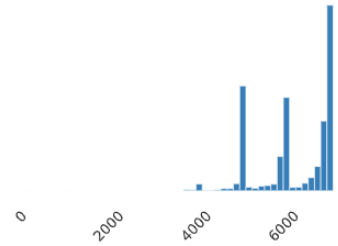
Distinct	7029	**Minimum**	0
Distinct (%)	20.1%	**Maximum**	9728
Missing	18	**Zeros**	3
Missing (%)	0.1%	**Zeros (%)**	< 0.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	2605.1349	**Memory size**	547.9 KiB



generation nuclear

Real number (ℝ)

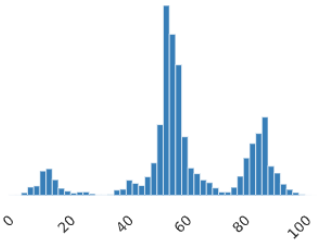
Distinct	2388	**Minimum**	0
Distinct (%)	6.8%	**Maximum**	7117
Missing	17	**Zeros**	3
Missing (%)	< 0.1%	**Zeros (%)**	< 0.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	6263.8833	**Memory size**	547.9 KiB



generation other

Real number (ℝ)

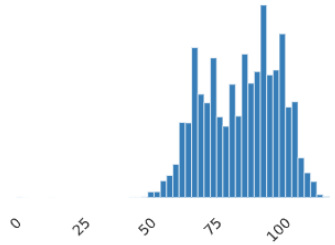
Distinct	103	**Minimum**	0
Distinct (%)	0.3%	**Maximum**	106
Missing	18	**Zeros**	4
Missing (%)	0.1%	**Zeros (%)**	< 0.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	60.229077	**Memory size**	547.9 KiB



generation other renewable

Real number (ℝ)

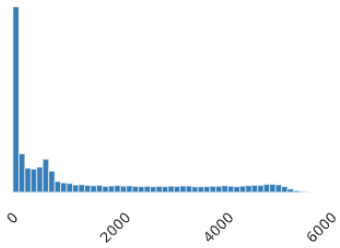
Distinct	78	**Minimum**	0
Distinct (%)	0.2%	**Maximum**	119
Missing	18	**Zeros**	3
Missing (%)	0.1%	**Zeros (%)**	< 0.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	85.640063	**Memory size**	547.9 KiB



generation solar

Real number (ℝ)

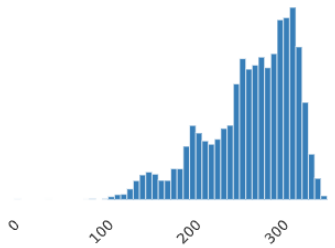
Distinct	5331	**Minimum**	0
Distinct (%)	15.2%	**Maximum**	5792
Missing	18	**Zeros**	3
Missing (%)	0.1%	**Zeros (%)**	< 0.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1432.7054	**Memory size**	547.9 KiB



generation waste

Real number (ℝ)

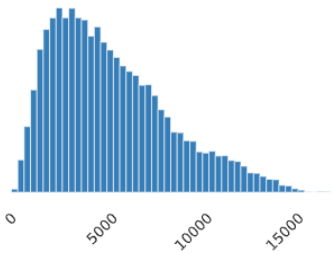
Distinct	262	**Minimum**	0
Distinct (%)	0.7%	**Maximum**	357
Missing	19	**Zeros**	3
Missing (%)	0.1%	**Zeros (%)**	< 0.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	269.45423	**Memory size**	547.9 KiB



generation wind onshore

Real number (ℝ)

Distinct	11465	**Minimum**	0
Distinct (%)	32.7%	**Maximum**	17436
Missing	18	**Zeros**	3
Missing (%)	0.1%	**Zeros (%)**	< 0.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	5464.4537	**Memory size**	547.9 KiB

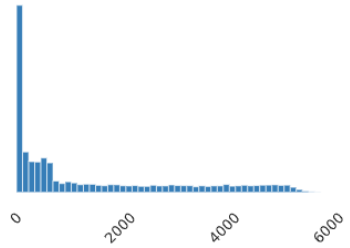


forecast solar day ahead

Real number (ℝ)

HIGH CORRELATION ZEROS

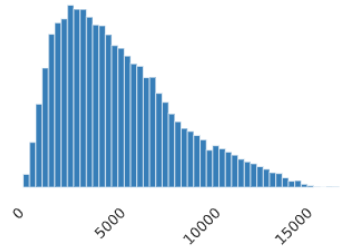
Distinct	5356	**Minimum**	0
Distinct (%)	15.3%	**Maximum**	5836
Missing	0	**Zeros**	539
Missing (%)	0.0%	**Zeros (%)**	1.5%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1439.1073	**Memory size**	547.9 KiB



forecast wind onshore day ahead

Real number (ℝ)

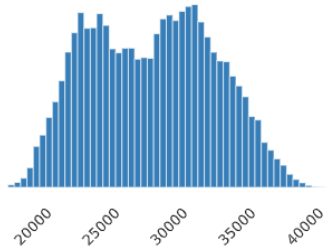
Distinct	11332	**Minimum**	237
Distinct (%)	32.3%	**Maximum**	17430
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	5471.1892	**Memory size**	547.9 KiB



total load forecast

Real number (ℝ)

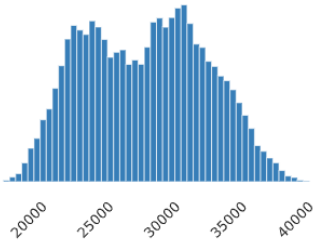
Distinct	14790	**Minimum**	18105
Distinct (%)	42.2%	**Maximum**	41390
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	28712.204	**Memory size**	547.9 KiB



total load actual

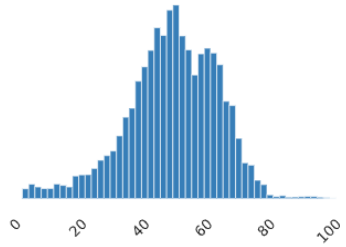
Real number (ℝ)

Distinct	15126	**Minimum**	18041
Distinct (%)	43.2%	**Maximum**	41015
Missing	36	**Zeros**	0
Missing (%)	0.1%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	28697.034	**Memory size**	547.9 KiB



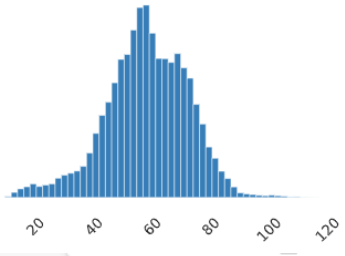
price day ahead
Real number (ℝ)

Distinct	5747	**Minimum**	2.06
Distinct (%)	16.4%	**Maximum**	101.99
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	49.874335	**Memory size**	547.9 KiB



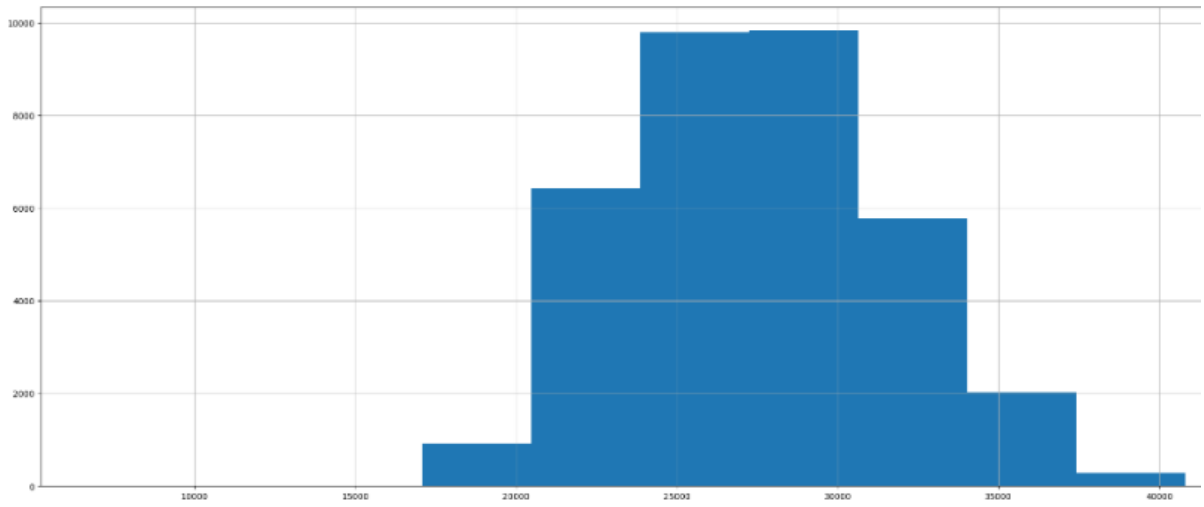
price actual
Real number (ℝ)

Distinct	6653	**Minimum**	9.33
Distinct (%)	19.0%	**Maximum**	116.8
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	57.883808	**Memory size**	547.9 KiB



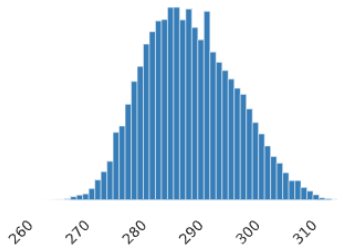

```
df1['total generation'].hist()
```

```
<Axes: >
```



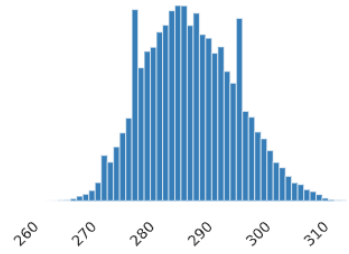
temp
Real number (ℝ)

Distinct	20742	**Minimum**	262.24
Distinct (%)	12.3%	**Maximum**	315.6
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	289.64778	**Memory size**	2.6 MiB



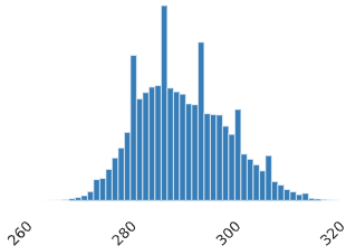
temp_min
Real number (ℝ)

Distinct	18552	**Minimum**	262.24
Distinct (%)	11.0%	**Maximum**	315.15
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	288.28813	**Memory size**	2.6 MiB



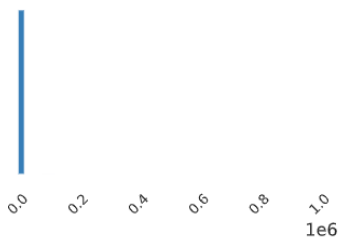
temp_max
Real number (ℝ)

Distinct	18590	**Minimum**	262.24
Distinct (%)	11.0%	**Maximum**	321.15
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	291.20216	**Memory size**	2.6 MiB



pressure
Real number (ℝ)

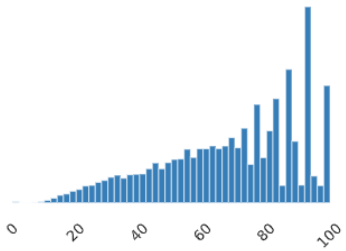
Distinct	190	**Minimum**	0
Distinct (%)	0.1%	**Maximum**	1008371
Missing	0	**Zeros**	2
Missing (%)	0.0%	**Zeros (%)**	< 0.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1073.1208	**Memory size**	2.6 MiB



humidity

Real number (ℝ)

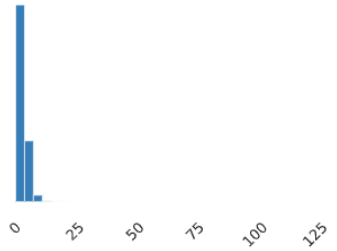
Distinct	100	**Minimum**	0
Distinct (%)	0.1%	**Maximum**	100
Missing	0	**Zeros**	60
Missing (%)	0.0%	**Zeros (%)**	< 0.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	68.10116	**Memory size**	2.6 MiB



wind_speed

Real number (ℝ)

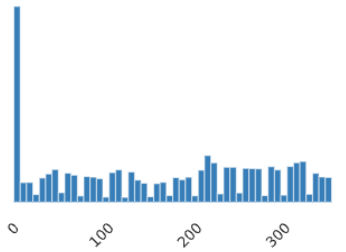
Distinct	36	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	133
Missing	0	**Zeros**	17639
Missing (%)	0.0%	**Zeros (%)**	10.5%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	2.4869165	**Memory size**	2.6 MiB



wind_deg

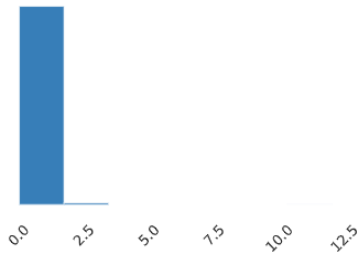
Real number (ℝ)

Distinct	361	**Minimum**	0
Distinct (%)	0.2%	**Maximum**	360
Missing	0	**Zeros**	24139
Missing (%)	0.0%	**Zeros (%)**	14.3%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	166.93122	**Memory size**	2.6 MiB



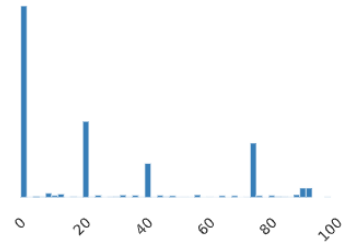
rain_1h
Real number (ℝ)

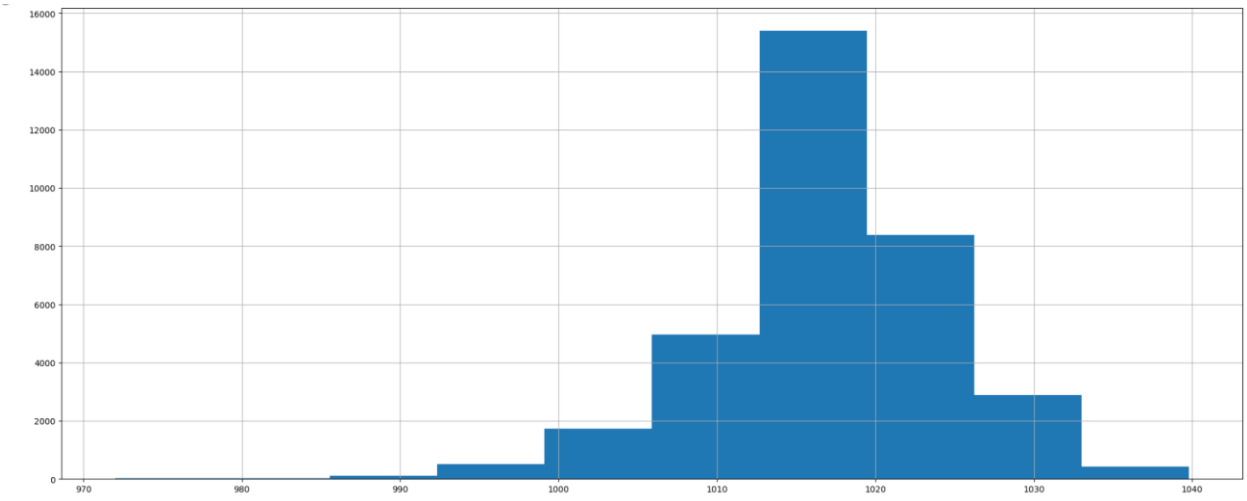
Distinct	7	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	12
Missing	0	**Zeros**	149986
Missing (%)	0.0%	**Zeros (%)**	89.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.077657905	**Memory size**	2.6 MiB



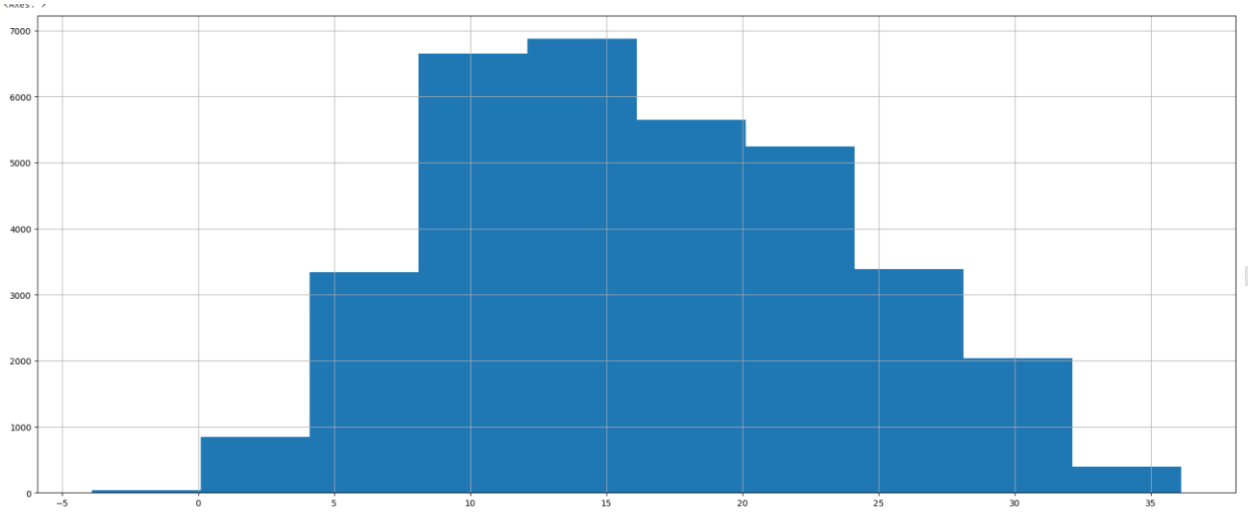
clouds_all
Real number (ℝ)

Distinct	97	**Minimum**	0
Distinct (%)	0.1%	**Maximum**	100
Missing	0	**Zeros**	76824
Missing (%)	0.0%	**Zeros (%)**	45.6%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	25.192371	**Memory size**	2.6 MiB

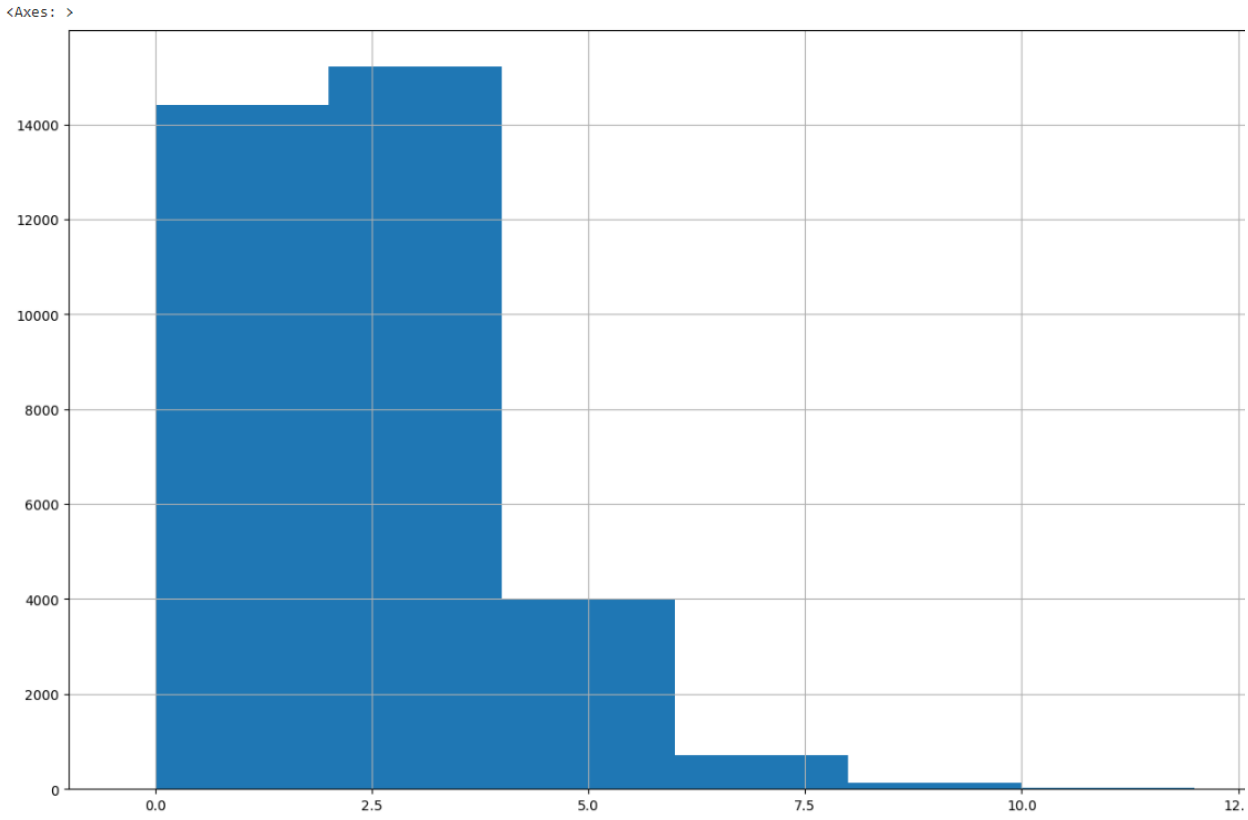




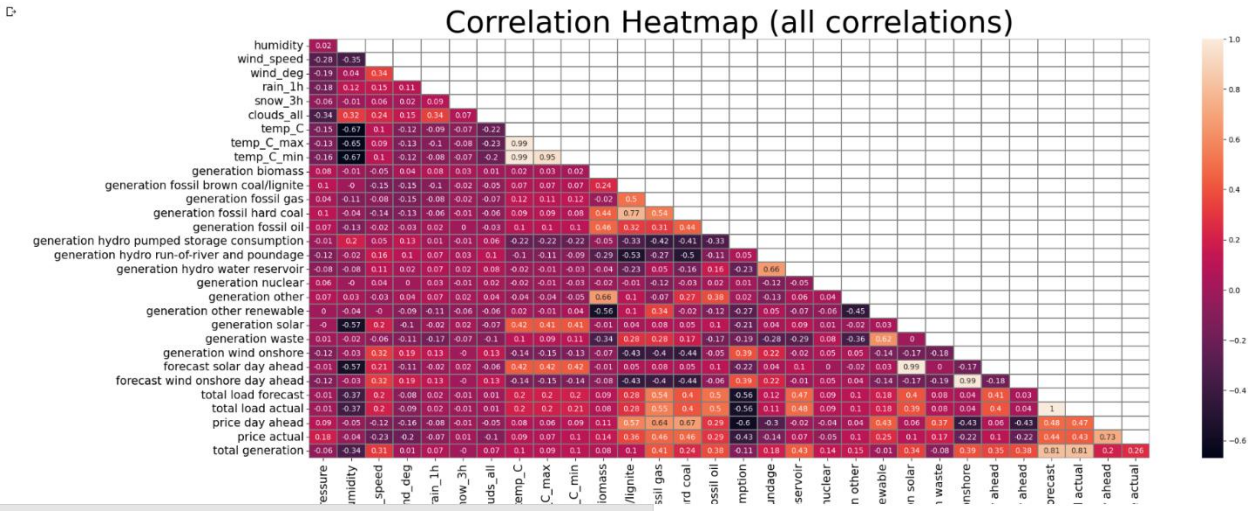
Pressure Histogram after treatment of extreme values



Temperature in degree Celsius Histogram



Wind_speed Histogram after treatment of extreme values




```
# Confusion Matrix for Logistic Regression
actual = y_test
predicted = log_prediction
Confusion_Matrix_Log_Reg = confusion_matrix(actual, predicted)
Confusion_Matrix_Log_Reg
```

```
array([[135,  11],
       [  8, 138]])
```

```
# Confusion Matrix for KNN Classifier
actual = y_test
predicted = KNN_prediction
Confusion_Matrix_KNN= confusion_matrix(actual, predicted)
Confusion_Matrix_KNN
```

```
array([[129,  17],
       [ 13, 133]])
```