ELEC-E8113

Information systems in industry

R teamwork

Author:	Version:	Date:
Ilkka Seilonen	1.0	5.9.2022

Table of contents

1 Introduction

In this R assignment you are supposed to learn about data collection and analysis in the information systems in industry though implementing a simple data analysis application for forecasting consumption of electric power. The assignment is divided into two parts:

1. Part one: Forecasting consumption of electric power with time-series models in two simple test cases based on sample data.
2. Part two: Implementation of a data collection application and updating the data for forecasting.

1.1 Part one: data analysis

An electric power distribution company wants to be able to forecast consumption of their customers. This is needed in order to be able to acquire proper amount of power from power plants and power markets. For forecasting it is good to note the characteristics of the behavior of the customers. Different types of customers consume electric power in different ways, e.g. apartment households vs. townhouses with possible electric power heating. There are also identifiable periods in consumption, e.g. daily, weekly and annually. The annual period is mainly due to electric heating and electric lighting. In this assignment you are going to forecast consumption of two example customers, identified with numbers one and two, based on their previous consumption and outside temperature.

The forecasting of the power distribution company requires time-series modeling of the consumption of electric power in the following way.

a) As a starting point you have a database with three tables in an SQL server (MariaDB, accessible through HeidiSQL, see chapter 2.3) describing two different consumers of electric power. The database contains past measurements of the consumption of electric power (table “cons”) and outside temperature at the locations of the consumers (table “temp”). In addition to this, there are forecasts of temperature for a given time period (table “tempfc”). Details about the data is described in chapter 1.3.

b) In order to create forecasts for both consumers, you first have to read the files to R (see chapter 2.4) for calculations and then write the result back to the database. You can first develop your R calculations interactively with RStudio and later run them in RStudio or through Rscrip (see chapter 2.4).

c) You can present the forecasts (and the original data if you like) with RStudio (see chapter 2.5). At this stage you should also assess how good your forecasts are and are there justifiable reasons to trust them.

1.2 Part two: data collection

Another activity of the power distribution company, in addition to data analysis, is data collection. The data to be analyzed has to be collected and updated from available sources. A data collection application that reads the required data, transforms it into a suitable form and stores it has to be designed and run on a regular base.

In this assignment the initial data collection has already been made. The data to be analyzed is already in the database. Your job is now to design a data collection application that could have collected this data and is able to update it.

The updating of the data required for forecasting will be done in the following way.

a) As a starting point you have six message queues in an AMQP server (accessible e.g. through a Python application, see chapter 2.6) containing messages about new measurements and forecasts at a later point of time. You need to read the data in the messages and write it to the database.

b) In order to perform the data update you need to create a Python application and run it (see chapter 2.6). The application has to read new messages, read their data and write it to the correct table in the database.

After you have updated the data in the database you can run the R scripts again and observe new forecasts with the Shiny applications.

1.3 Existing data

1.3.1 Input data

The existing electricity consumption and temperature measurements and forecasts are in three tables in the database: cons, temp and tempfc. You can study the database through HeidiSQL (see chapter 2.3). The contents of the tables are explained below.

The table “cons” describes electricity consumption of customers. Each row describes a measurement with the following information. About the customer “1” you have measurements for a time period 1.1. - 30.09.2012 and about the customer “2” for 1.1.2013 – 15.3.2013. Note that all rows do not necessary contain complete information.

- date (yyyy-mm-dd)
- electricity consumption (kWh)
- time (1-24) which means the number of the measurement within a day (“1” has only two measurements in a day whereas “2” has 24)
- site (1 or 2) which identifies the customer “1” and “2”

The table “temp” describes temperature measurements for customers. Each row describes a measurement in a similar way than in the previous case. Also the time periods of the measurements are the same. The data in each row is explained below. Again, note that all rows do not necessary contain all information.

- date (yyyy-mm-dd)
- temperature (C)
- site (1 or 2) similarly to “cons” table
- time (1-24) similarly to “cons” table (customer “1” has only one measurement in a day)

The table “tempfc” is similar to “temp”. However, it describes temperature forecasts, not measurements.

1.3.2 Update data

The updated electricity consumption and temperature measurements and forecasts are assumed to be available in six message queues in an AMQP server: cons1, temp1, tempfc1, cons2, temp2 and tempfc2. You can imagine that you have received these messages through web services. First you need to write messages to the queues. Then you can study the contents of the message queues with tools of the RabbitMQ server (see chapter 2.6). The messages contain similar information than the database but for different time periods. You can study the contents of the messages yourself.

1.3.3 Result data

The database contains also a table called “confc”. This is the electricity consumption forecast you are requested to calculate. The table indicates what information will be needed in the resulting forecasts. The data is explained below.

- forecast (kWh)
- lo80, hi80, lo95 and hi95 (kWh) are confidence intervals reported by the forecast function of R
- id (> 1) is the number of the row in this forecast (this is for the user interface application)
- site (1 or 2) similarly to “cons” and “temp” tables
- time (1-24) similarly to “cons” and “temp” tables
- date (yyyy-mm-dd)

2 Implementation tools

2.1 Overview

This exercise has to be done with the server e8113-1.org.aalto.fi. You can access the server using VPN or from the computers in the room AS5 (computer class) in Maarintie 8. Just go to the room, log in with your Aalto account and use the tools explained in the following chapters.

2.2 SSH and SCP

You can make an SSH connection to the server with a command similar to the following. Replace “team00” with your own team name.

```
ssh -oHostKeyAlgorithms=+ssh-dss team00@e8113-1.org.aalto.fi
```

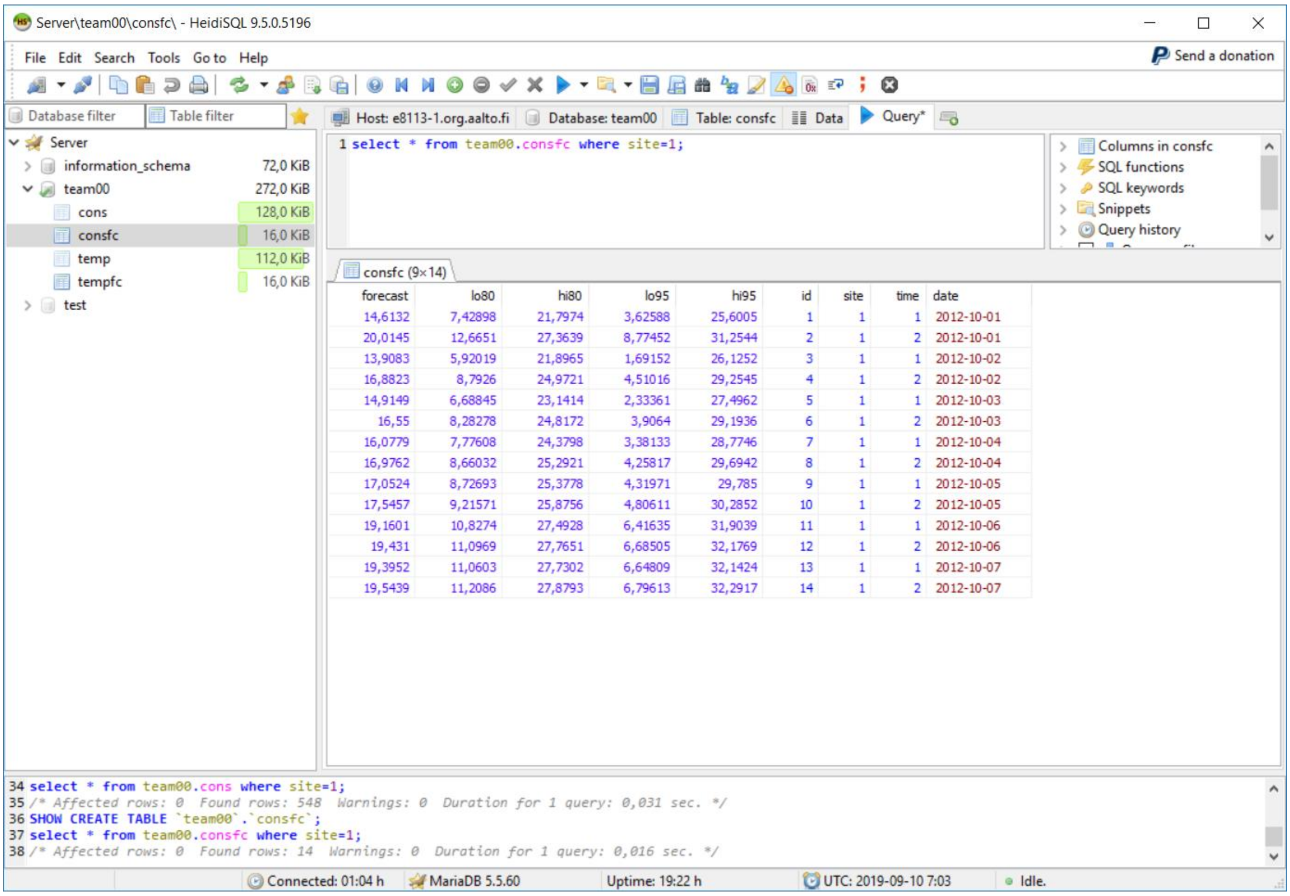
You can also copy files to and from the server with scp. An example command is below.

```
scp -oHostKeyAlgorithms=+ssh-dss hh_update.py team00@e8113-1.org.aalto.fi:/home/team00
```

2.3 HeidiSQL

HeidiSQL is a client application for accessing SQL servers. HeidiSQL lets you manage the data stored in an SQL server. You can study the data in the tables of your database. You can also modify and delete data from the database if needed. However, be extremely careful if deleting data from the database! A good way to handle the data is SQL queries. If you need instructions for this check SQL documentation in the Internet.

In order to use HeidiSQL you can install its portable version available in Mycourses. Unzip the zip file and run heidisql.exe. Then connect to e8113-1.org.aalto.fi with your team’s username and password. The port number is 3306. You should be able to see your data at the SQL server (see Figure 1). If you need instructions for using HeidiSQL check its documentation in the Internet.



```
select * from team00.consfc where site=1;
```

forecast	lo80	hi80	lo95	hi95	id	site	time	date
14,6132	7,42898	21,7974	3,62588	25,6005	1	1	1	2012-10-01
20,0145	12,6651	27,3639	8,77452	31,2544	2	1	2	2012-10-01
13,9083	5,92019	21,8965	1,69152	26,1252	3	1	1	2012-10-02
16,8823	8,7926	24,9721	4,51016	29,2545	4	1	2	2012-10-02
14,9149	6,68845	23,1414	2,33361	27,4962	5	1	1	2012-10-03
16,55	8,28278	24,8172	3,9064	29,1936	6	1	2	2012-10-03
16,0779	7,77608	24,3798	3,38133	28,7746	7	1	1	2012-10-04
16,9762	8,66032	25,2921	4,25817	29,6942	8	1	2	2012-10-04
17,0524	8,72693	25,3778	4,31971	29,785	9	1	1	2012-10-05
17,5457	9,21571	25,8756	4,80611	30,2852	10	1	2	2012-10-05
19,1601	10,8274	27,4928	6,41635	31,9039	11	1	1	2012-10-06
19,431	11,0969	27,7651	6,68505	32,1769	12	1	2	2012-10-06
19,3952	11,0603	27,7302	6,64809	32,1424	13	1	1	2012-10-07
19,5439	11,2086	27,8793	6,79613	32,2917	14	1	2	2012-10-07

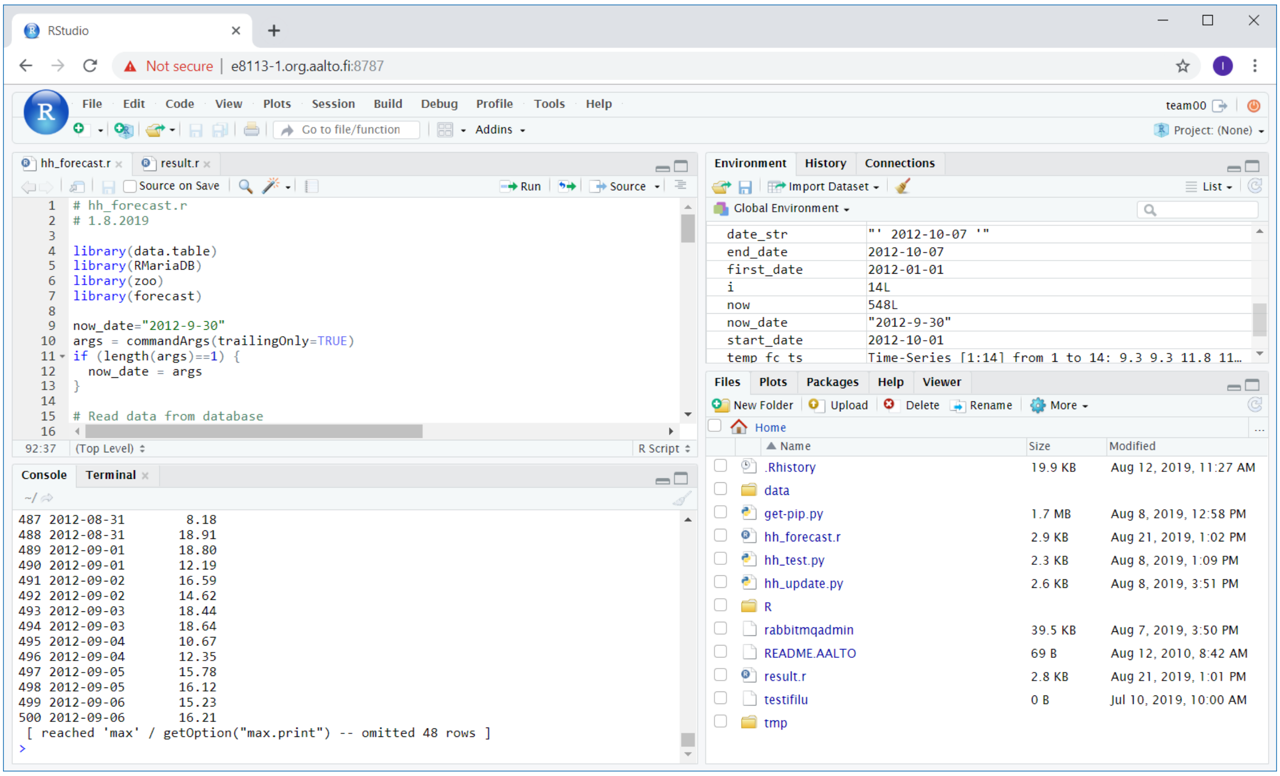
```
select * from team00.cons where site=1;
/* Affected rows: 0  Found rows: 548  Warnings: 0  Duration for 1 query: 0,031 sec. */
SHOW CREATE TABLE `team00`.`consfc`;
select * from team00.consfc where site=1;
/* Affected rows: 0  Found rows: 14  Warnings: 0  Duration for 1 query: 0,016 sec. */
```

Figure 1. HeidiSQL user interface.

2.4 RStudio

R is a software for statistical computing. In this assignment you are first going to utilize RStudio for developing your application and then using Rscript to run it. Your applications need to be saved as R source files in your home directory at the server. RStudio is a tool for developing your application interactively. Rscript you can call from a command shell through an SSH connection. In your application you are supposed to read input data from the SQL database and write your results there as well.

You can use RStudio with a browser at the address *http://e8113-1.org.aalto.fi:8787* (see Figure 2). Log in with your team name and password. If you need instructions for using RStudio, check the documentation provided by the R project and other material available in the Internet. In your home directory there should be a skeleton R scripts *forecast1.r* or *forecast2.r* that provides a few useful hints.



```
# hh_forecast.r
# 1.8.2019

library(data.table)
library(RMariaDB)
library(zoo)
library(forecast)

now_date="2012-9-30"
args = commandArgs(trailingOnly=TRUE)
if (length(args)==1) {
  now_date = args
}

# Read data from database

```

```
487 2012-08-31        8.18
488 2012-08-31       18.91
489 2012-09-01       18.80
490 2012-09-01       12.19
491 2012-09-02       16.59
492 2012-09-02       14.62
493 2012-09-03       18.44
494 2012-09-03       18.64
495 2012-09-04       10.67
496 2012-09-04       12.35
497 2012-09-05       15.78
498 2012-09-05       16.12
499 2012-09-06       15.23
500 2012-09-06       16.21
 [ reached 'max' / getOption("max.print") -- omitted 48 rows ]
>
```

date_str	"' 2012-10-07 '"
end_date	2012-10-07
first_date	2012-01-01
i	14L
now	548L
now_date	"2012-9-30"
start_date	2012-10-01
temp_fc_ts	Time-Series [1:14] from 1 to 14: 9.3 9.3 11.8 11…

▲ Name	Size	Modified
.Rhistory	19.9 KB	Aug 12, 2019, 11:27 AM
data		
get-pip.py	1.7 MB	Aug 8, 2019, 12:58 PM
hh_forecast.r	2.9 KB	Aug 21, 2019, 1:02 PM
hh_test.py	2.3 KB	Aug 8, 2019, 1:09 PM
hh_update.py	2.6 KB	Aug 8, 2019, 3:51 PM
R		
rabbitmqadmin	39.5 KB	Aug 7, 2019, 3:50 PM
README.AALTO	69 B	Aug 12, 2010, 8:42 AM
result.r	2.8 KB	Aug 21, 2019, 1:01 PM
testifilu	0 B	Jul 10, 2019, 10:00 AM
tmp		

Figure 2. User interface of RStudio.

2.5 Viewing data and forecasts

R has several graphical functions for viewing data and forecasts. It is very useful to use them for understanding the original data and forecasts. For example, observations may have trends, cycles and various amount of noise. Forecasts have reliability associated to them. You can view the data and forecasts with the following commands in RStudio.

plot(cons_ts)

where cons_ts is the consumption data converted to time series type with function *ts()*, and

autoplot(cons_fc)

where cons_fc is a consumption forecast calculated with function *forecast()*.

See the RStudio documentation in the Internet for other ways to view data and forecasts.

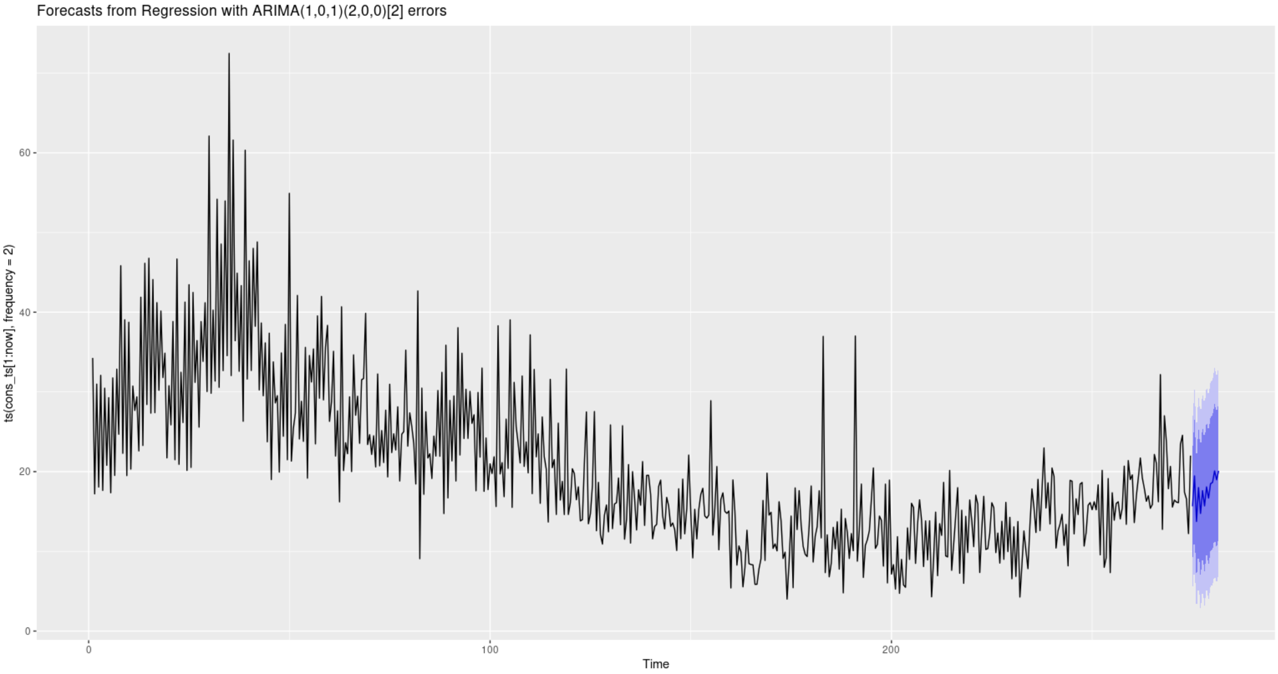


Figure 3. Forecast for customer "1".

2.6 RabbitMQ

Whereas RStudio is utilized for analyzing data stored in an SQL database the purpose of RabbitMQ is to buffer collected data before it is written to the database. RabbitMQ is an implementation of a message queue following the AMQP specification. In this assignment the RabbitMQ server simulates a message queue storing data received from various data acquisition services. You can receive new measurements and forecasts through it.

You can access messages in RabbitMQ using scripts from a command shell (see examples below). For the assignment you have to implement a Python script that reads the messages from the RabbitMQ server and writes their data to the SQL server. In your home directory there should be a skeleton scripts *update1.py* and *update2.py* that provides a few useful hints.

You can access messages in the RabbitMQ server using a Python script called *rabbimqadmin* with commands similar to the following. Replace "team00" with your own team name and password and other parameters accordingly. You can write messages to the server, list contents of queues and remove the messages. When writing messages are read from files in the folder */home/share*.

```
python rabbitmqadmin --host=localhost --vhost=team00 --user=team00 --password=xxxxxx list queues
```

```
python rabbitmqadmin --host=localhost --vhost=team00 --user=team00 --password=xxxxxx purge queue name=tempfc1
```

```
cat /home/share/tempfc1_2012-11-01.json | while read line; do python rabbitmqadmin --host=localhost --vhost=team00 --user=team00 --password=xxxxxx publish routing_key=tempfc1 payload="$line"; done
```

3 Useful links

The following links are likely to contain useful information about time-series modelling, R and RabbitMQ. You can find more through Google searches. However, note that all information that you may find might not be relevant to those versions of R and RabbitMQ you have, or might not be relevant at all. Be critical about the information you find.

Time series analysis

MS-C2128 – Prediction and Time Series Analysis (year 2018). Particularly see the slides of lectures 4 – 6.

30E00800 - Time Series Analysis. Particularly see the slides of chapter 5 of the course book (http://www.cambridge.org/features/economics/brooks/PPT.html).

R, RStudio

https://www.r-project.org/

https://cran.r-project.org/manuals.html

https://www.rdocumentation.org/

https://www.rstudio.com

SQL

https://www.w3schools.com/sql/

RabbitMQ

https://www.rabbitmq.com/

https://pika.readthedocs.io/en/stable/intro.html

4 Deliverables

You need to demonstrate the teacher that your application works and you have a justified understanding how good your forecasts are. You will need to present arguments about how good your forecasts are and to which extent one can trust them. You are requested to provide the following deliverables:

1. Programs. You need to present the R and Python source code of your applications that are able to create the requested forecasts and update the data as needed.

2. Demonstration. You have to agree a demonstration event with the teacher. In the event you are requested to create forecasts three times for both customers in three different situations in the following way:

   - Situation 1. Create forecasts from data in the database in the beginning. For the customer “1” the date of the last measurements is 2012-09-30 and the length of the requested forecast period seven days. For the customer “2” is date is 2013-03-15 and forecasting period three days.
   - Situation 2. Update measurements and forecasts in the database with the values from messages in the message queues and create forecasts again similarly to situation 1. Now the dates of last measurements are 2012-10-31 and 2013-03-31 and the lengths of the forecasting periods the same (i.e. seven and three days).
   - Situation 3. Update measurements and forecasts in the database again with the new set of messages in the message queues and create forecasts again similarly to previous situations. You are only going to receive the new messages in the demonstration event, not before. The dates of last measurements will be 2012-11-30 and 2013-04-15.

3. Document. You need to create a short (preferably no more than 5 pages) document about the (mathematical) forecasting method in your application. You need to describe the identified ARIMA models and their estimated parameters and assess their error terms and the reliability of the forecasts. In addition to this, you need to assess the forecasts from the viewpoint of the phenomenon itself, i.e. consumption of electric power by ordinary consumers. Would the forecasts make sense? This document can be written after the demonstration event.

Appendix A: FAQ

Q1: We have serious trouble and cannot make any progress. What should we do?

A1: Check the course web page on mycouses.aalto.fi first. If that does not help then send email to ilkka.seilonen@aalto.fi**.**

Q2: Is the electricity consumption data seasoable?

A2: Good question. You are supposed to find it out!

In general time series can contain seasonability, i.e. pattern that keeps occuring at regular times. You can tell auto.arima and ts functions if your data contains seasonability. You study your data e.g. with the following methods:

https://anomaly.io/seasonal-trend-decomposition-in-r/index.html

Or you can try using seasonability and see if you get better results.

Usually customers use electricity more during day than night.

Q3: Why should I care about the error terms or residuals after model fitting?

A3: The end results of the teamwork are forecasts with certain reliability. The reliability is often described with "prediction intervals" (e.g. 80% and 95%). These calculated with the R forecast function. There is a connection from residuals (in model fitting) to the reliability. See e.g. the following page:

https://otexts.com/fpp2/prediction-intervals.html

Q4: What is the minimum set of results needed for accepted teamwork?

A4: You have to be able to demonstrate creation of forecasts in at least three situations. This means that you will need a working data update function, too. In addition, you need to provide the required document and sources of your code.

Q5: Can I ask the teacher if my answer is “right” or “perfect”?

A5: The teacher should refuse to answer such questions. The quality of your answer will be evaluated after the teamwork. You should have a reason to believe that your answer is “good enough”.

The teacher is supposed to provide hints about how to make progress and overcome problematic situations.