Data Science Methods in Finance

Take Home Exam

Name	SNR	email
Khalid Amine	2045967	k.amine@tilburguniversity.edu

Table of Contents

- Question 1: Report summary graphs for the data
- Question 2: Prepare the dataset for making daily price predictions
- Question 3: Predicting electricity prices in the first hour of the day
- Question 4: Predicting hourly electricity prices
- Question 5: Trading strategy

Question 1: Report summary graphs for the data

Question 1.1 (1)

Average hourly prices over a day, across countries

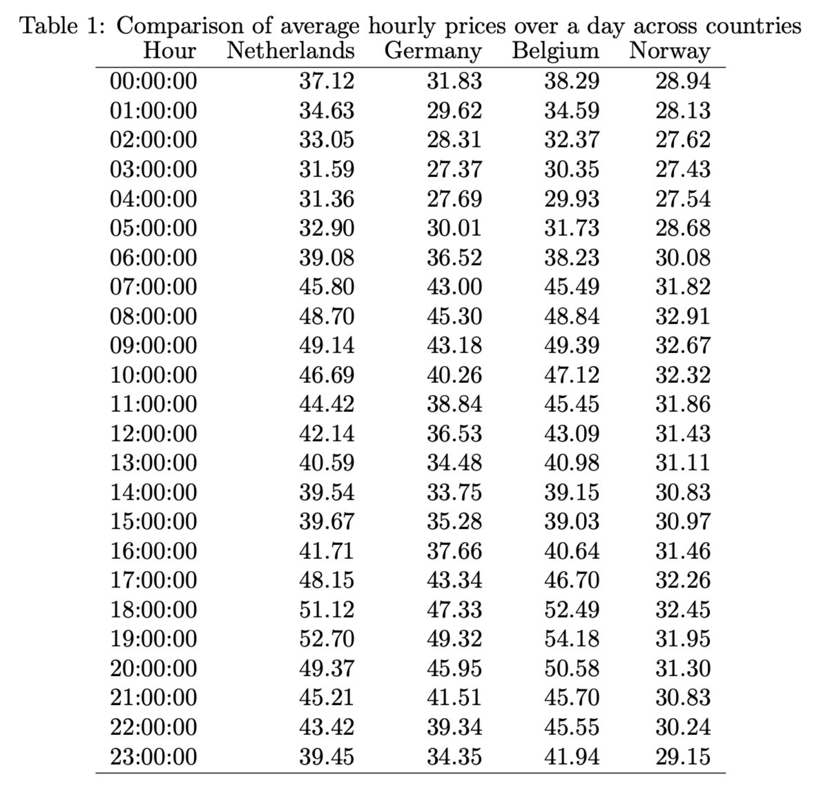
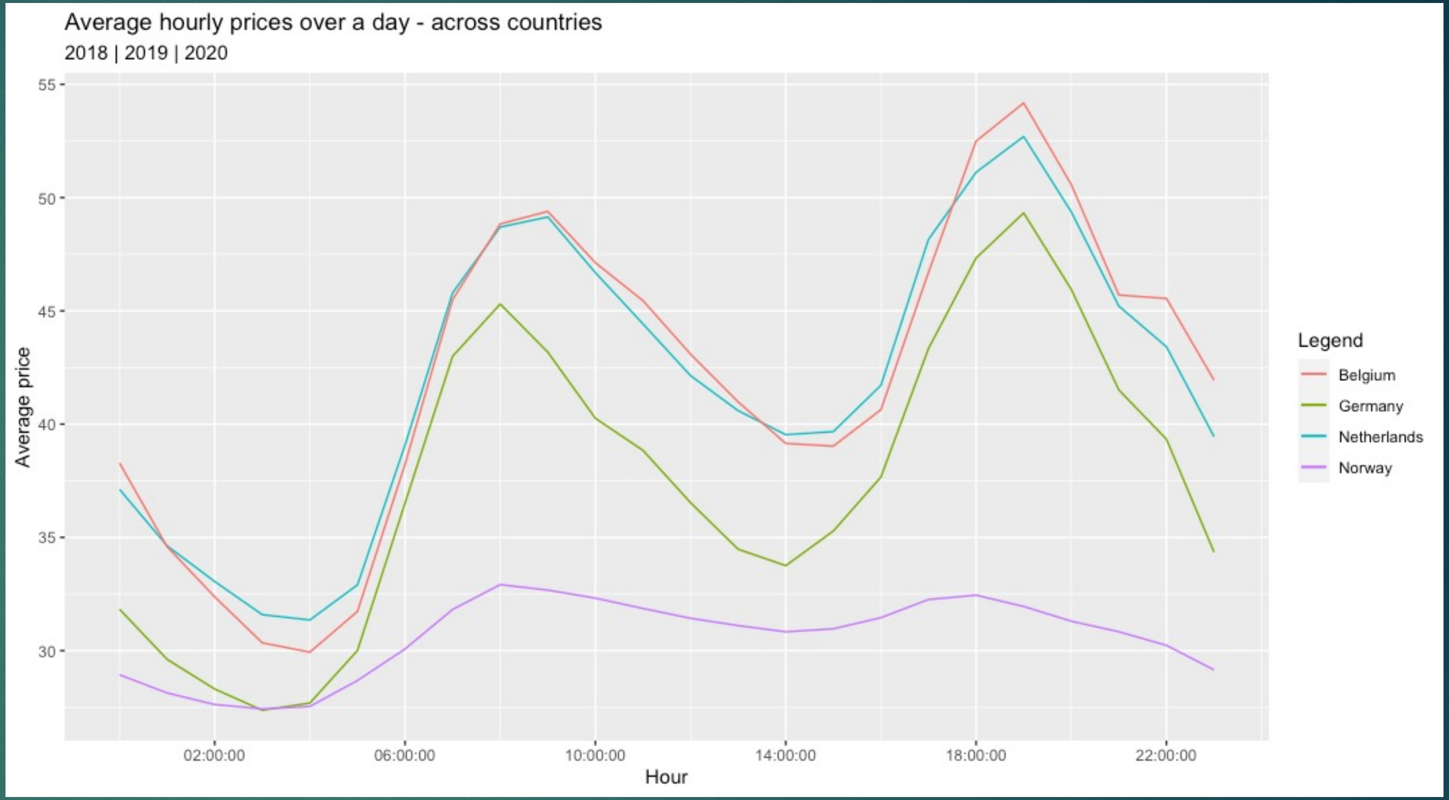


Table 1: Comparison of average hourly prices over a day across countries

Hour	Netherlands	Germany	Belgium	Norway
00:00:00	37.12	31.83	38.29	28.94
01:00:00	34.63	29.62	34.59	28.13
02:00:00	33.05	28.31	32.37	27.62
03:00:00	31.59	27.37	30.35	27.43
04:00:00	31.36	27.69	29.93	27.54
05:00:00	32.90	30.01	31.73	28.68
06:00:00	39.08	36.52	38.23	30.08
07:00:00	45.80	43.00	45.49	31.82
08:00:00	48.70	45.30	48.84	32.91
09:00:00	49.14	43.18	49.39	32.67
10:00:00	46.69	40.26	47.12	32.32
11:00:00	44.42	38.84	45.45	31.86
12:00:00	42.14	36.53	43.09	31.43
13:00:00	40.59	34.48	40.98	31.11
14:00:00	39.54	33.75	39.15	30.83
15:00:00	39.67	35.28	39.03	30.97
16:00:00	41.71	37.66	40.64	31.46
17:00:00	48.15	43.34	46.70	32.26
18:00:00	51.12	47.33	52.49	32.45
19:00:00	52.70	49.32	54.18	31.95
20:00:00	49.37	45.95	50.58	31.30
21:00:00	45.21	41.51	45.70	30.83
22:00:00	43.42	39.34	45.55	30.24
23:00:00	39.45	34.35	41.94	29.15



Average hourly prices over a day - across countries

2018 | 2019 | 2020

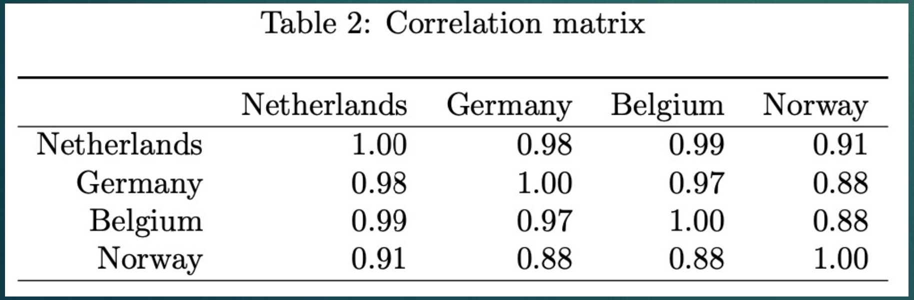


Table 2: Correlation matrix

	Netherlands	Germany	Belgium	Norway
Netherlands	1.00	0.98	0.99	0.91
Germany	0.98	1.00	0.97	0.88
Belgium	0.99	0.97	1.00	0.88
Norway	0.91	0.88	0.88	1.00

- From the price plot, it seems that the electricity price of Belgium and Germany are almost perfectly identical to that of the Netherlands. This is confirmed by the very high (almost identical) correlations with these two countries, as shown in table 2.

Question 1.1 (2)

Average hourly prices over a day in the Netherlands (full week against only weekend)

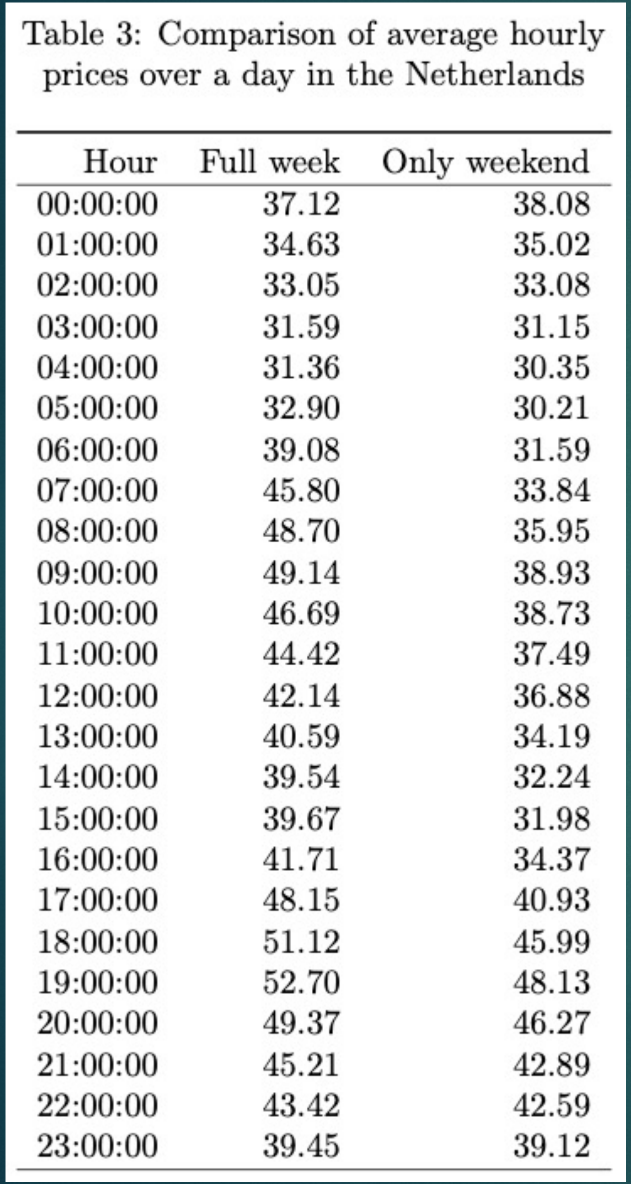
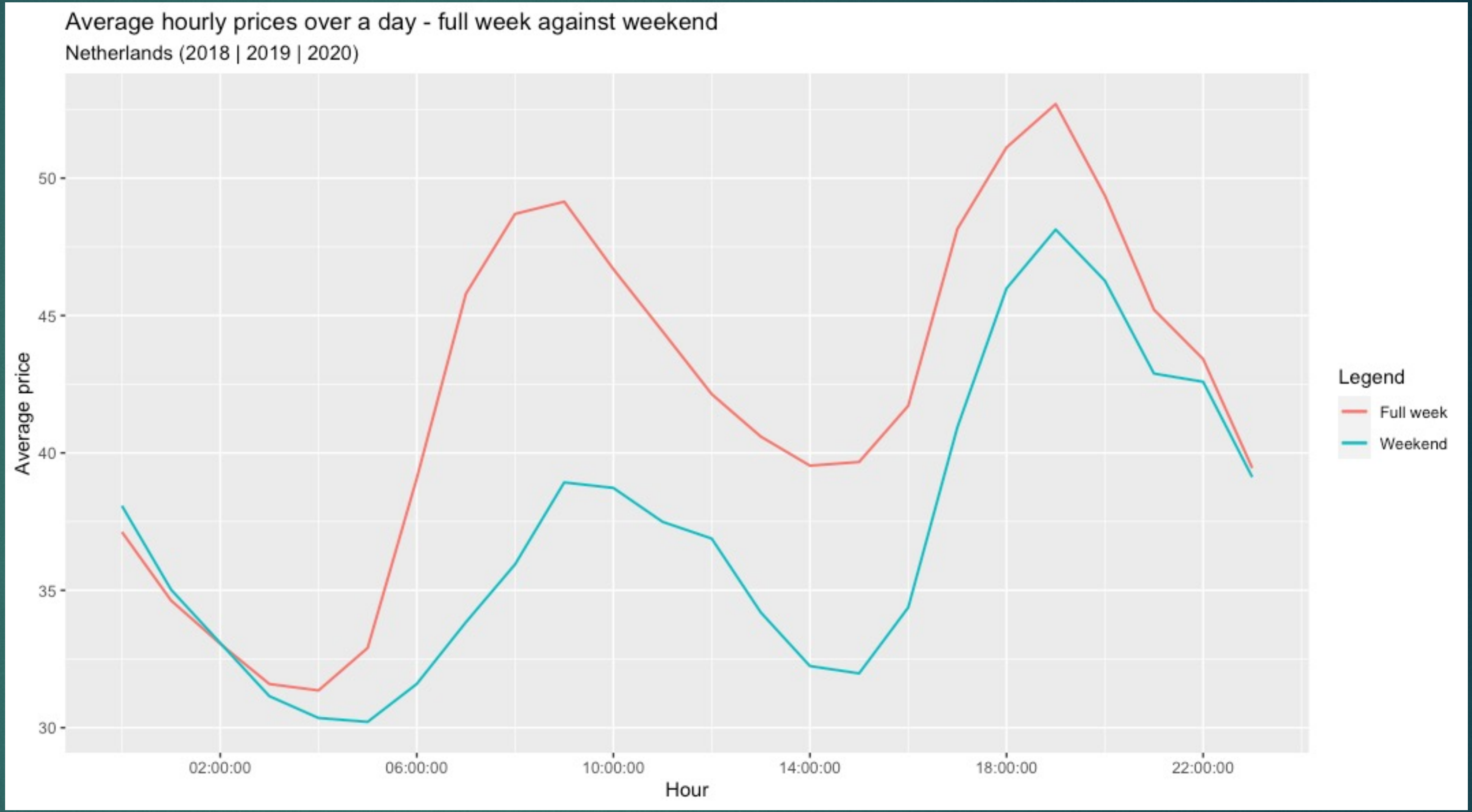


Table 3: Comparison of average hourly prices over a day in the Netherlands

Hour	Full week	Only weekend
00:00:00	37.12	38.08
01:00:00	34.63	35.02
02:00:00	33.05	33.08
03:00:00	31.59	31.15
04:00:00	31.36	30.35
05:00:00	32.90	30.21
06:00:00	39.08	31.59
07:00:00	45.80	33.84
08:00:00	48.70	35.95
09:00:00	49.14	38.93
10:00:00	46.69	38.73
11:00:00	44.42	37.49
12:00:00	42.14	36.88
13:00:00	40.59	34.19
14:00:00	39.54	32.24
15:00:00	39.67	31.98
16:00:00	41.71	34.37
17:00:00	48.15	40.93
18:00:00	51.12	45.99
19:00:00	52.70	48.13
20:00:00	49.37	46.27
21:00:00	45.21	42.89
22:00:00	43.42	42.59
23:00:00	39.45	39.12



- In the Netherlands, it seems that the prices during the weekends are on average lower. An explanation for this could be that during the weekend, much fewer factories and businesses are active which creates less demand for electricity. By the law of supply and demand, the decrease in demand should lower the prices, ceteris paribus.

Question 1.2

Volatility of hourly prices over a day in the Netherlands (full week against only weekend)

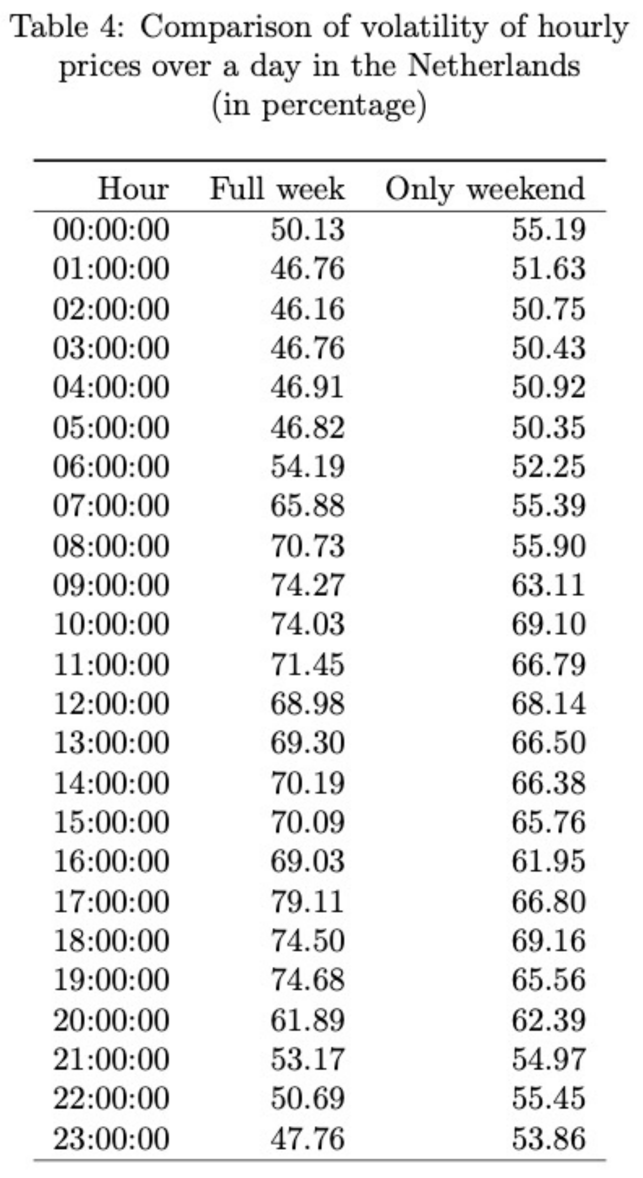
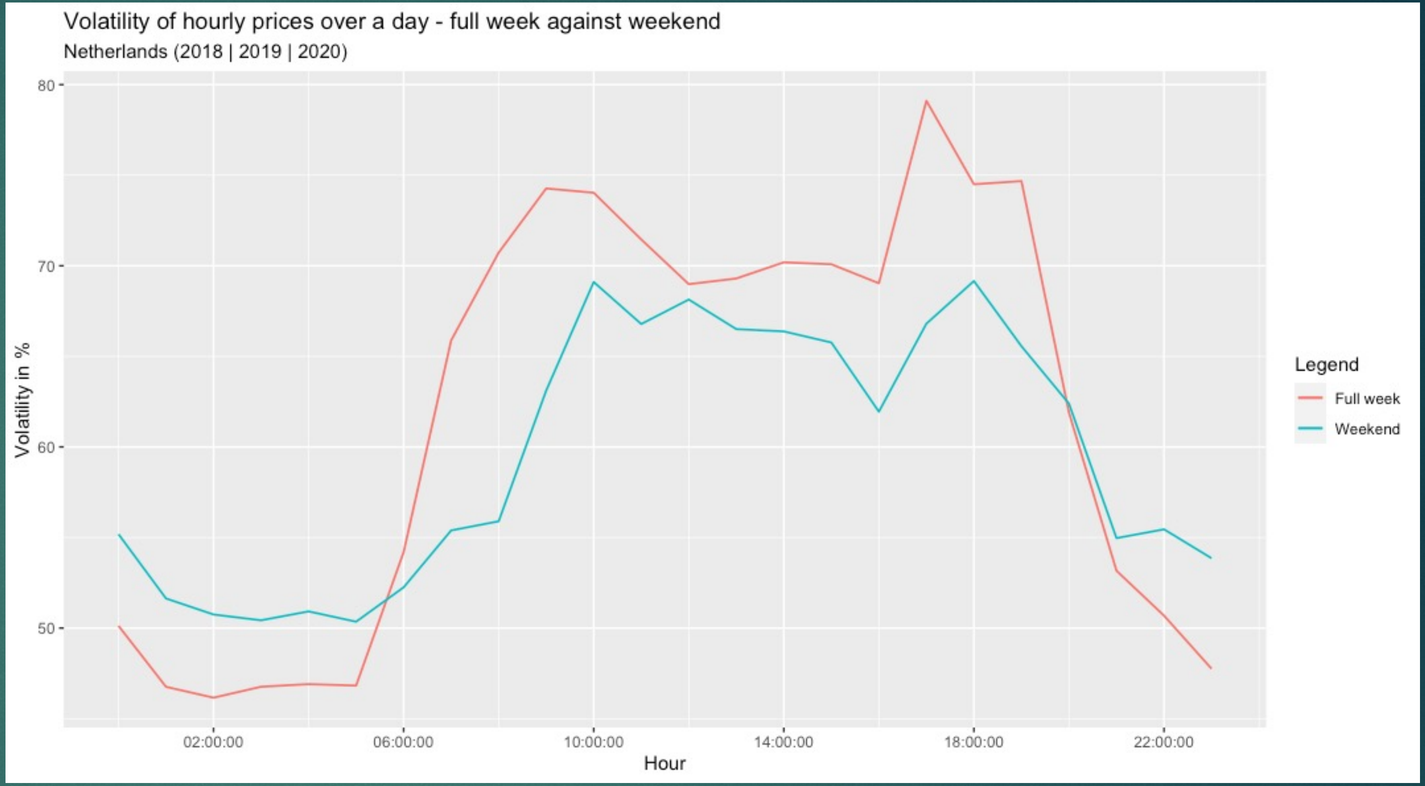


Table 4: Comparison of volatility of hourly prices over a day in the Netherlands (in percentage)

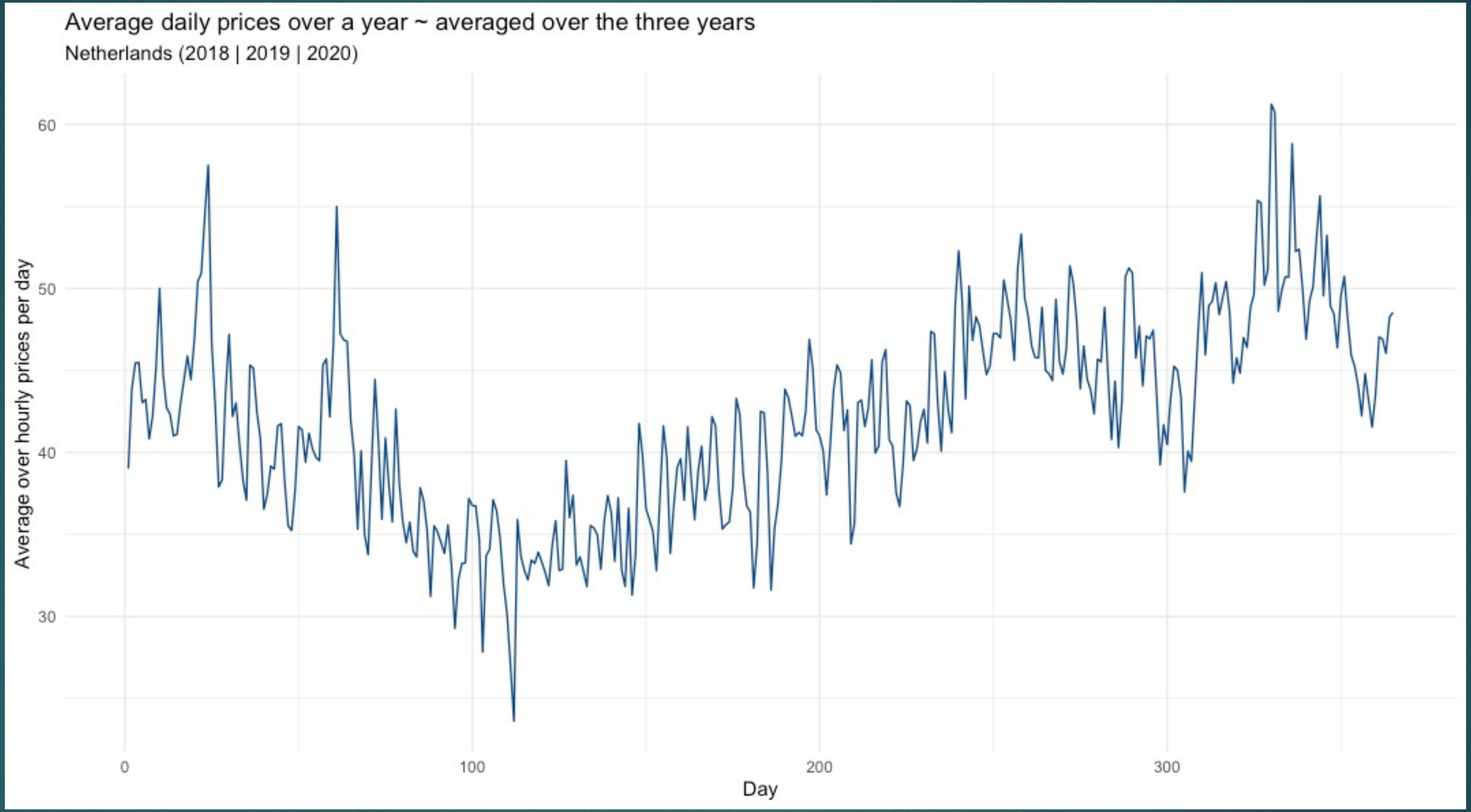
Hour	Full week	Only weekend
00:00:00	50.13	55.19
01:00:00	46.76	51.63
02:00:00	46.16	50.75
03:00:00	46.76	50.43
04:00:00	46.91	50.92
05:00:00	46.82	50.35
06:00:00	54.19	52.25
07:00:00	65.88	55.39
08:00:00	70.73	55.90
09:00:00	74.27	63.11
10:00:00	74.03	69.10
11:00:00	71.45	66.79
12:00:00	68.98	68.14
13:00:00	69.30	66.50
14:00:00	70.19	66.38
15:00:00	70.09	65.76
16:00:00	69.03	61.95
17:00:00	79.11	66.80
18:00:00	74.50	69.16
19:00:00	74.68	65.56
20:00:00	61.89	62.39
21:00:00	53.17	54.97
22:00:00	50.69	55.45
23:00:00	47.76	53.86



- Both lines are at its highest during the hours whereby people are on average awake (peak hours). The rapidly increasing demand during this time period increases the volatility.
- During the weekend there is on average less demand since factories and businesses are mostly closed, thus slower increases in demand during peak hours creates less volatility. It also seems like during the weekend, people tend to stay awake longer and also sleep longer, which is reflected by a less wavy line.

Question 1.3

Average daily prices over a year, averaged over the three years (2018, 2019, 2020)



- It seems like that during the colder periods of the year, the price increases. By the law of supply and demand, this can be explained by either an increase in demand or decrease in supply, ceteris paribus. An increase in demand seems to be the most obvious explanation, since during these periods additional energy will have to be used to heat businesses and homes.

Question 1.4

Average hourly solar generation over a day, April-September against October-March

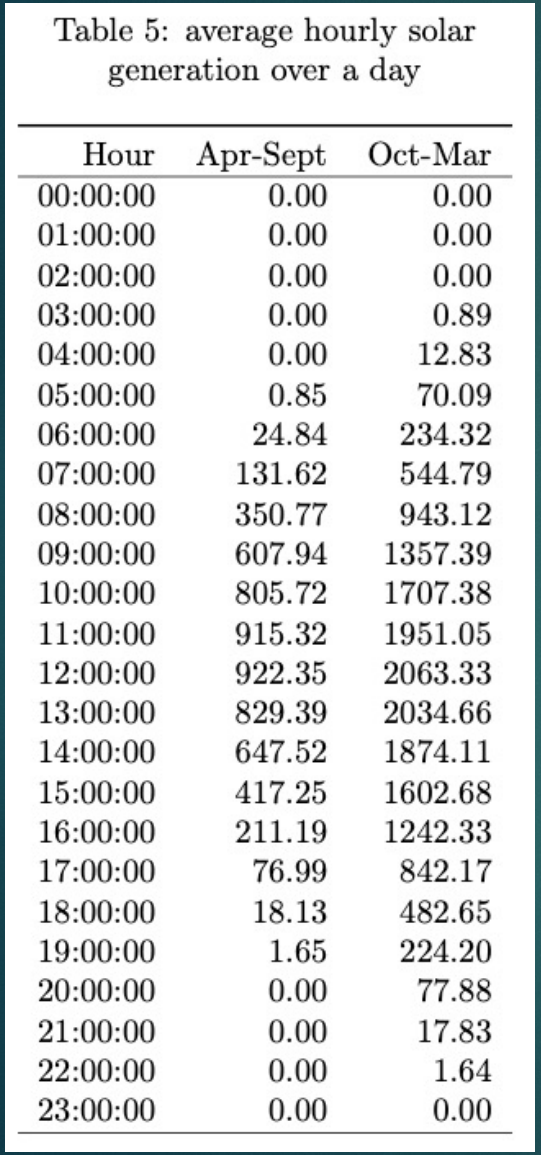
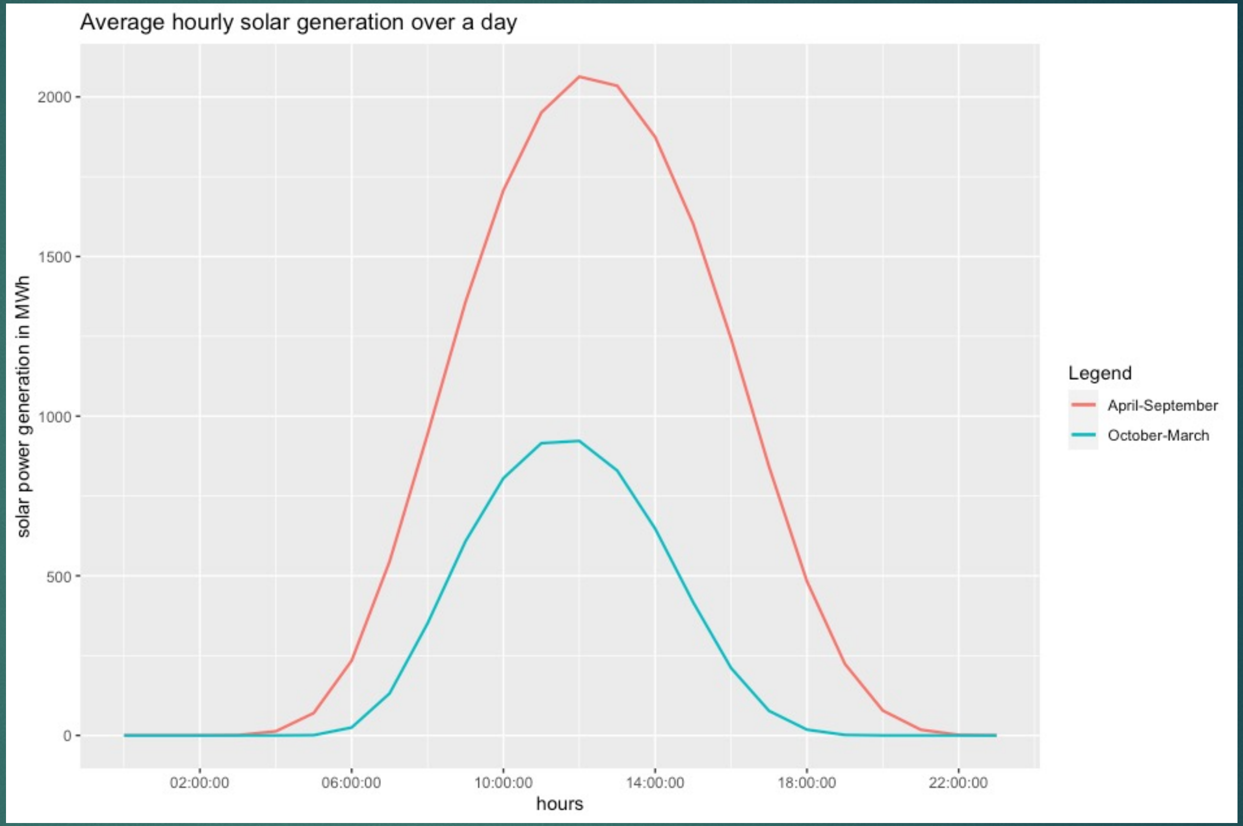


Table 5: average hourly solar generation over a day

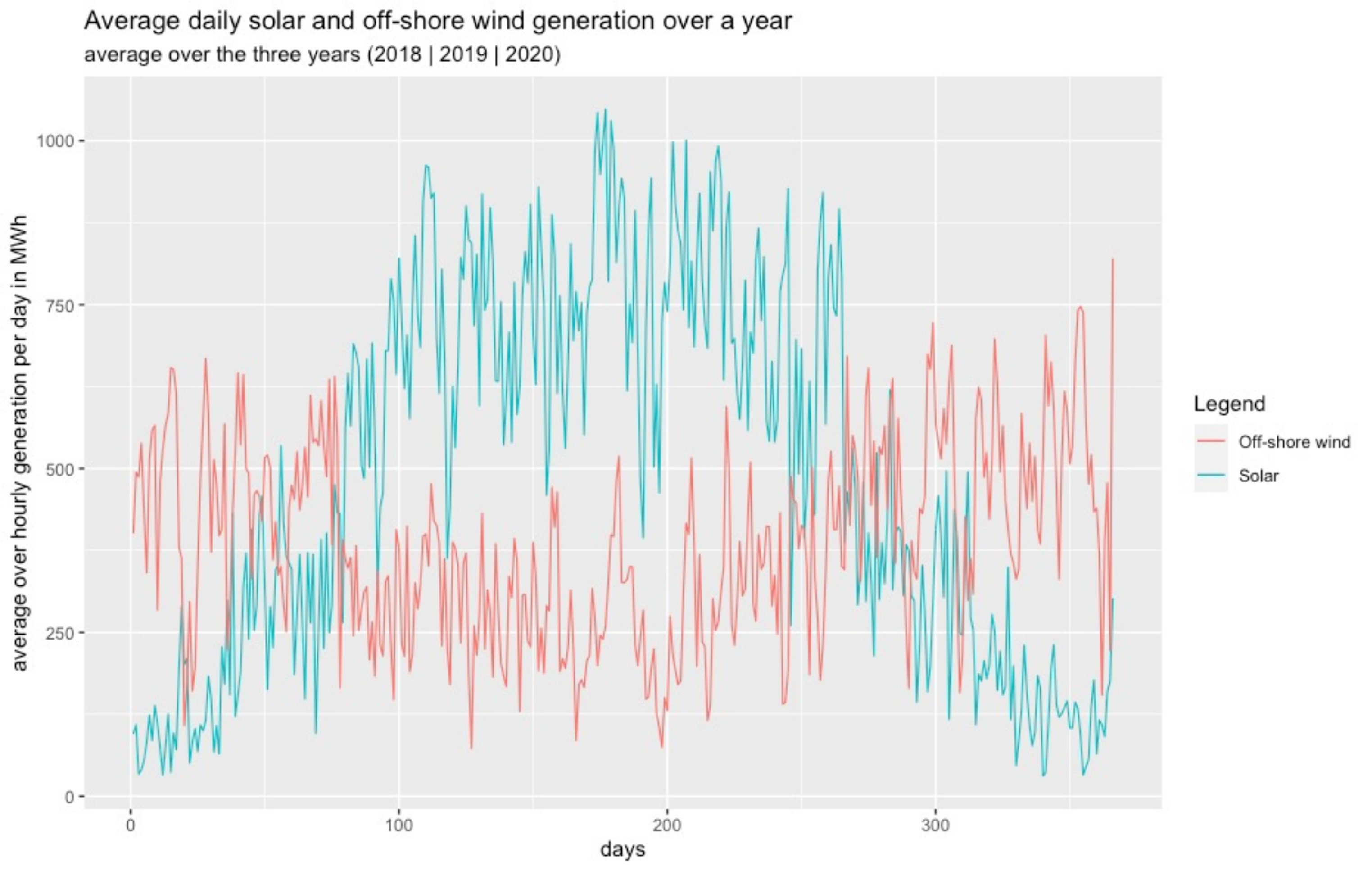
Hour	Apr-Sept	Oct-Mar
00:00:00	0.00	0.00
01:00:00	0.00	0.00
02:00:00	0.00	0.00
03:00:00	0.00	0.89
04:00:00	0.00	12.83
05:00:00	0.85	70.09
06:00:00	24.84	234.32
07:00:00	131.62	544.79
08:00:00	350.77	943.12
09:00:00	607.94	1357.39
10:00:00	805.72	1707.38
11:00:00	915.32	1951.05
12:00:00	922.35	2063.33
13:00:00	829.39	2034.66
14:00:00	647.52	1874.11
15:00:00	417.25	1602.68
16:00:00	211.19	1242.33
17:00:00	76.99	842.17
18:00:00	18.13	482.65
19:00:00	1.65	224.20
20:00:00	0.00	77.88
21:00:00	0.00	17.83
22:00:00	0.00	1.64
23:00:00	0.00	0.00



- It can clearly be inferred from the graph that during the sunnier period of the year (Apr-Sept), much more solar energy can be generated since the sun is stronger during these months.
- The width of the graph of the months Apr-Sept shows that during these months the earth's axis is tilted towards the sun's ecliptic, which not only causes warm weather but also longer hours of sunlight (Northern Hemisphere) which also results in more solar energy generation.

Question 1.5

Average daily solar and off-shore wind generation over a year



- During the summer, there is less wind compared to the winter which causes the graph to ”dip” in the middle of the year. Meaning that less energy is generated by off-shore wind during that period.
- As stated in question 1.4, during the summer the solar panels are exposed to more sunlight, which generates more solar energy during this period.

Question 2: Prepare the dataset for making daily price predictions

Question 2.1

Preparing the data

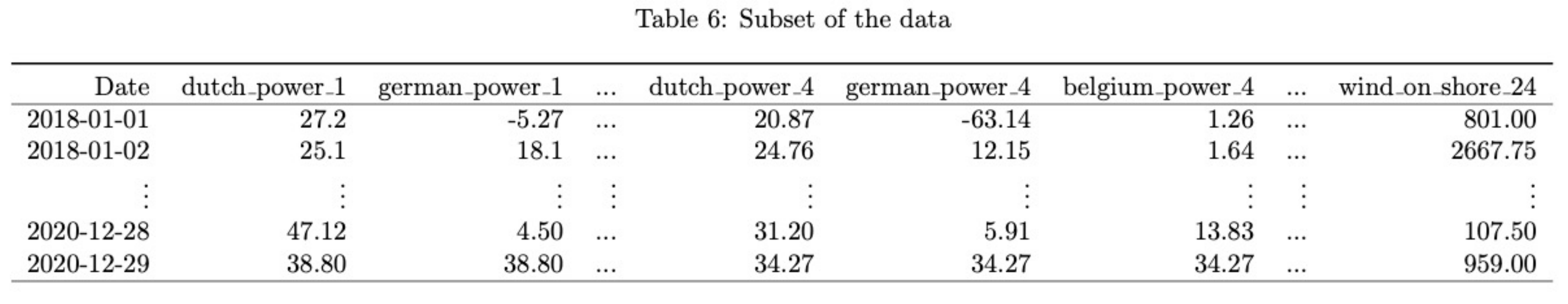


Table 6: Subset of the data

Date	dutch_power_1	german_power_1	...	dutch_power_4	german_power_4	belgium_power_4	...	wind_on_shore_24
2018-01-01	27.2	-5.27	...	20.87	-63.14	1.26	...	801.00
2018-01-02	25.1	18.1	...	24.76	12.15	1.64	...	2667.75
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2020-12-28	47.12	4.50	...	31.20	5.91	13.83	...	107.50
2020-12-29	38.80	38.80	...	34.27	34.27	34.27	...	959.00

Question 2.3

Clean data

1) Getting rid of the features with too many 0 entries

- Cut-off used: 15%, if a feature has more than 15% zero entries then this feature is removed.
- Removed features: solar_1_lag, solar_2_lag, solar_3_lag, solar_4_lag, solar_5_lag, solar_6_lag, solar_7_lag, solar_18_lag, solar_19_lag, solar_20_lag, solar_21_lag, solar_22_lag, solar_23_lag, solar_24_lag.
- Motivation for cut-off value of 15%: I used a low cut-off value of 15% in order to retain as much as possible, while still removing features that have very little to no impact. Most of the features removed are the solar features, since during those hours the sun has very little to no impact.
- Number of features in dataset left = 202

2) Splitting the data into training and test set: training set 2.5 years and test set 0.5 years

- 2.5 years is approximately 918 days and 0.5 years is approximately 182 days. Since the test set will be the smallest set, I first allocate precisely the last 182 days to test set and the remaining data (913 days) is then for the training set.
  - Training set: 2018-01-02 -> 2020-07-02 | Test set: 2020-07-03 -> 2020-12-29

3) Normalizing the data

- I normalized the data using the scale() function on the features (training and test set separately), such that all features have mean 0 and a standard deviation of 1.

Question 3: Predicting electricity prices in the first hour of the day

Question 3.1

Building a benchmark model ~ Predicting Dutch electricity price in the first hour of the day using the price in the last hour of the previous day (and a constant) as a predictor.

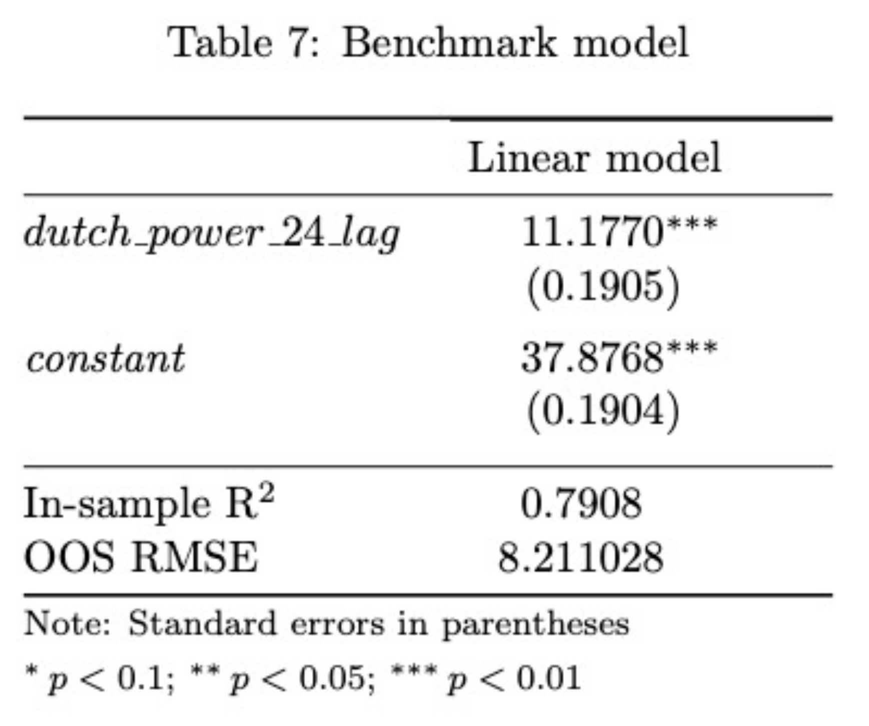


Table 7: Benchmark model

	Linear model
dutch_power_24_lag	11.1770***
	(0.1905)
constant	37.8768***
	(0.1904)
In-sample R^2	0.7908
OOS RMSE	8.211028

Note: Standard errors in parentheses

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

- The linear model seems to explain the variation very well in-sample, making it look like a very good model for predictions. The out-of-sample RMSE is rather high, higher than the ML models which will be discussed in the next slides.

Question 3.2 (1)

Machine learning models ~ Using different machine learning models to predict the electricity price in the first hour of the day.

- For these models, I use all 202 features which were the features left after removing the solar features that did not meet the required cut-off value of 15%.
  - Motivation for using these features: Having too many features may lead to overfitting, but this can be avoided by tuning the hyperparameters on the validation set. The more controlled information provided to a machine learning model, the better. Also, the Partial Least Squares model benefits from having as much as possible features since it aims to optimize the components that are related to the response variable.
  - Note: optimal parameters are chosen based on validation set performance (highest R-squared).

Question 3.2 (2)

Lasso model

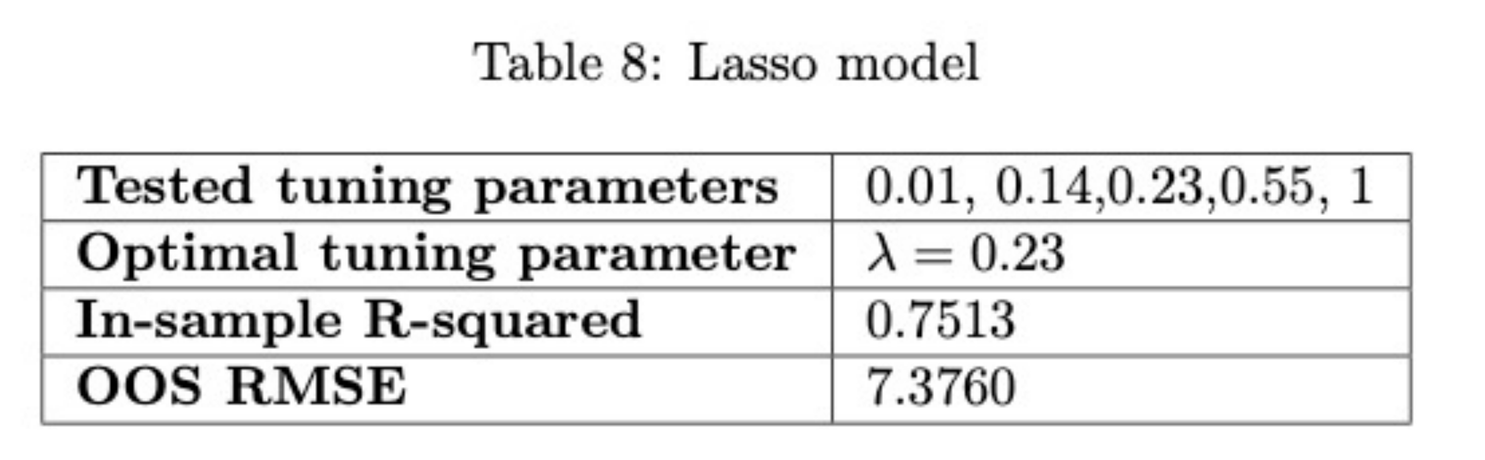


Table 8: Lasso model

Tested tuning parameters	0.01, 0.14,0.23,0.55, 1
Optimal tuning parameter	$\lambda = 0.23$
In-sample R-squared	0.7513
OOS RMSE	7.3760

- The Lasso regression seems to not overfit too much on the test data, since it has a moderately high R-squared. The out-of-sample RMSE is lower than that of the benchmark, beating the linear model in terms of out-of-sample performance meaning that the Lasso model has a better bias-variance trade-off.

Question 3.2 (3)

Partial Least Squares model

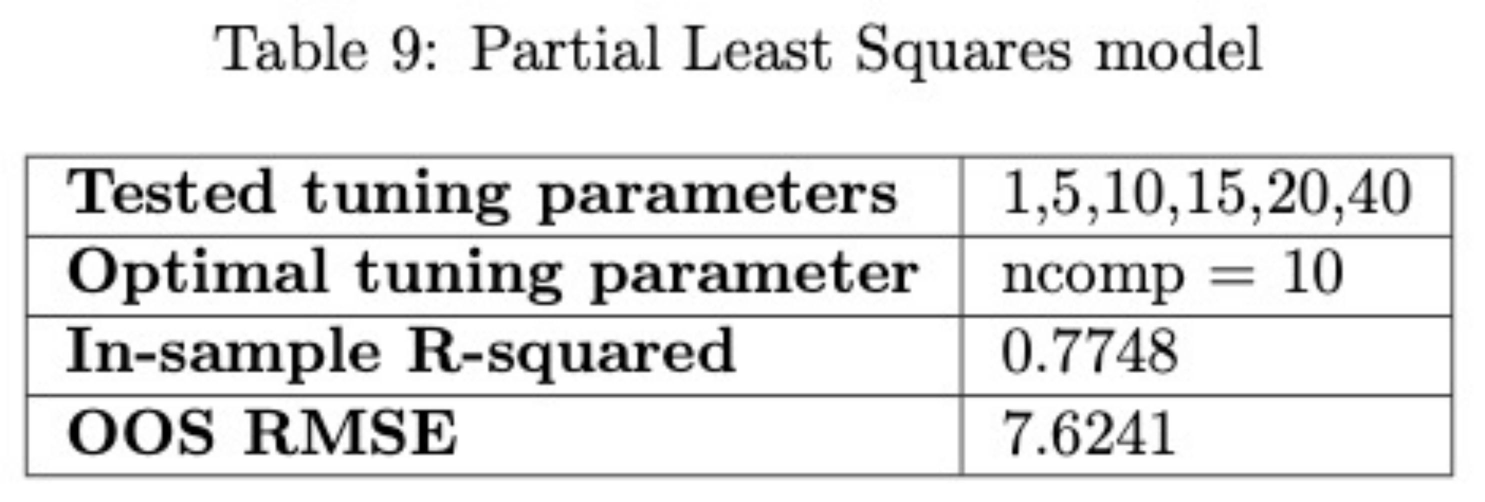


Table 9: Partial Least Squares model

Tested tuning parameters	1,5,10,15,20,40
Optimal tuning parameter	ncomp = 10
In-sample R-squared	0.7748
OOS RMSE	7.6241

- When compared to the Lasso, the PLS model tends to overfit a bit more on the training sample, since this model has a higher R-squared in-sample. This then resulted in a higher out-of-sample RMSE, which means that the lasso model is thus superior in terms of out-of-sample predictions.

Question 3.2 (4)

Random Forest model

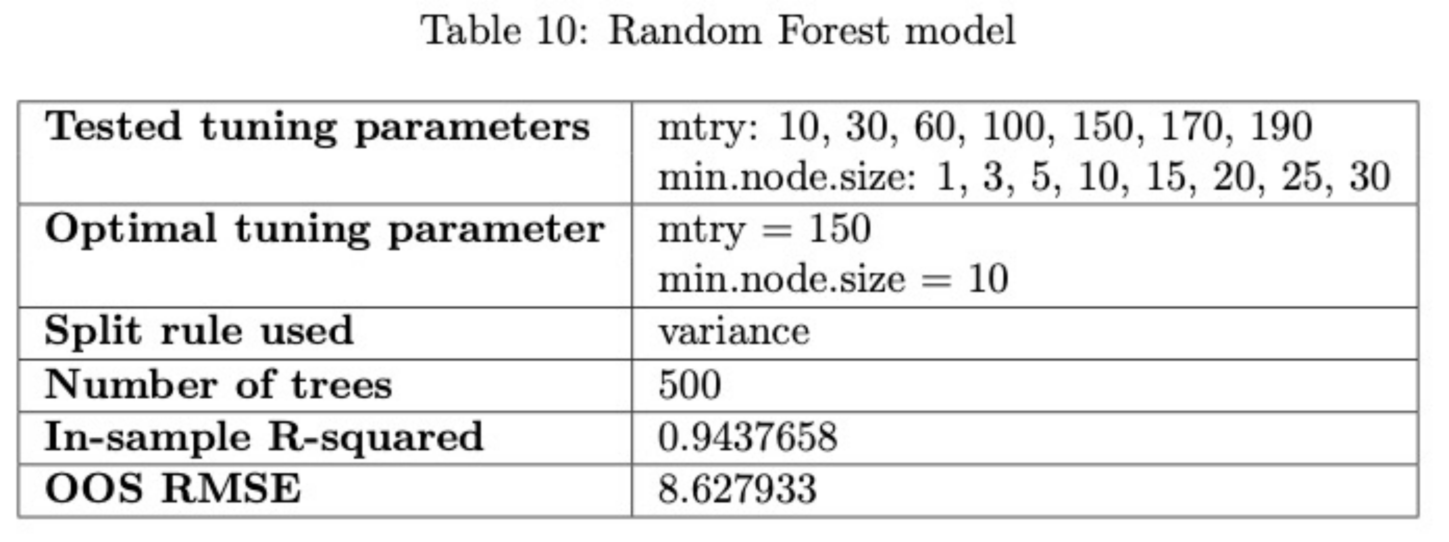


Table 10: Random Forest model

Tested tuning parameters	mtry: 10, 30, 60, 100, 150, 170, 190 min.node.size: 1, 3, 5, 10, 15, 20, 25, 30

Optimal tuning parameter	mtry = 150
	min.node.size = 10

Split rule used	variance
Number of trees	500
In-sample R-squared	0.9437658
OOS RMSE	8.627933

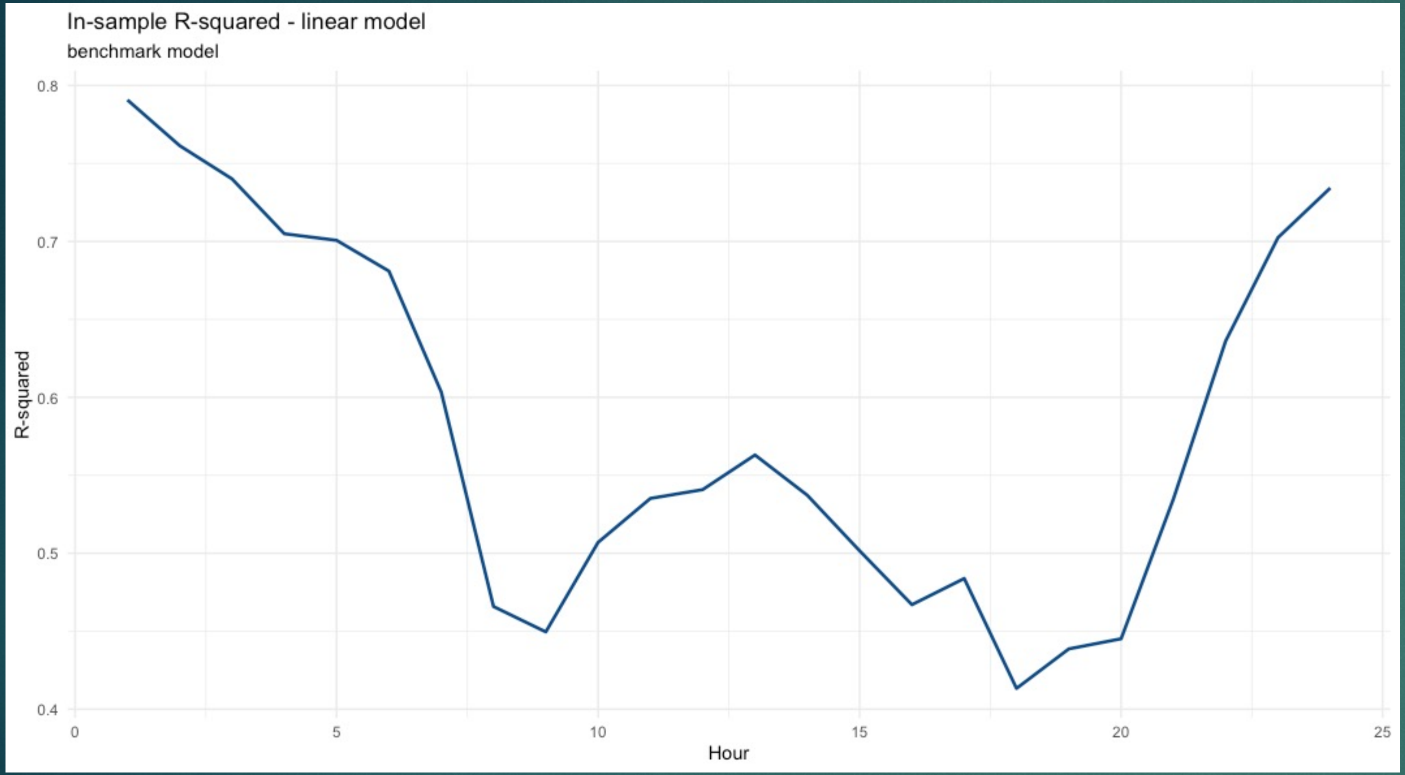
- The Random Forest model overfits heavily on the training set, despite tuning the hyperparameters on the validation set. This then resulted in a much higher out-of-sample RMSE compared to the other models, which means that the lasso model is thus far superior in terms of OOS predictions.

Question 4: Predicting hourly electricity prices

Question 4.1 (1)

Building a benchmark model ~ using 24 linear regressions to predict the 24 hourly Dutch electricity prices over the day using the price in the last hour of the previous day (and a constant) as a predictor.

- **In-sample R-squared plot**



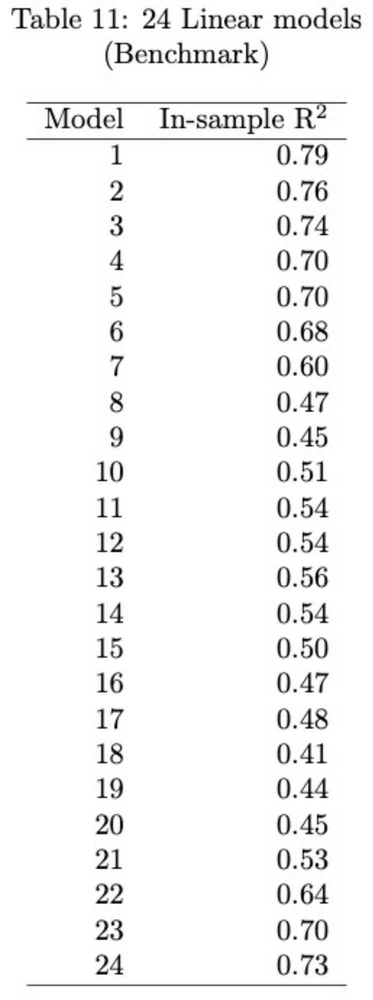


Table 11: 24 Linear models (Benchmark)

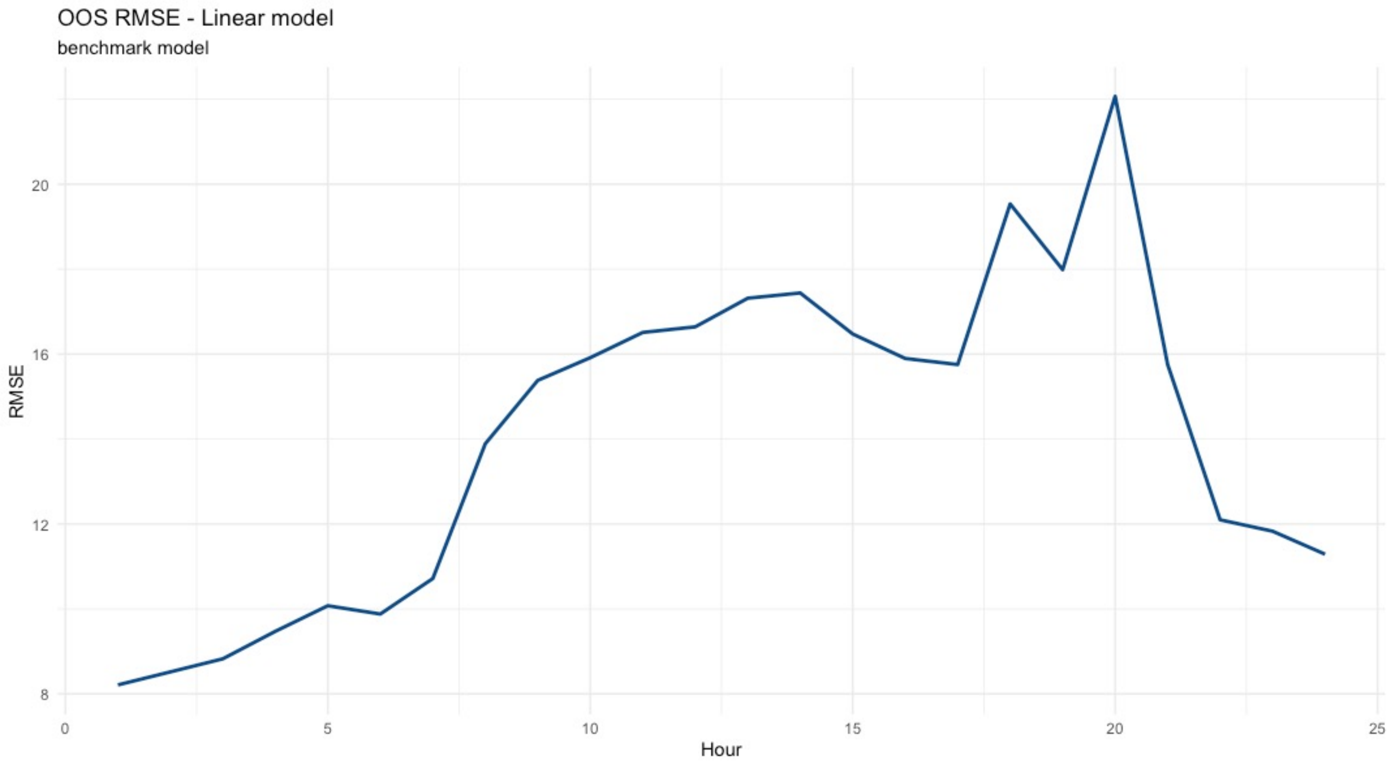
Model	In-sample R^2
1	0.79
2	0.76
3	0.74
4	0.70
5	0.70
6	0.68
7	0.60
8	0.47
9	0.45
10	0.51
11	0.54
12	0.54
13	0.56
14	0.54
15	0.50
16	0.47
17	0.48
18	0.41
19	0.44
20	0.45
21	0.53
22	0.64
23	0.70
24	0.73

- It seems that our linear model’s in-sample predictive power suffers a lot from the heavy increase in volatility during the peak hours. As inferred in question 1.2, in these hours the sharp increases in demand increases the volatility which makes it more difficult to predict future prices.

Question 4.1 (2)

Building a benchmark model ~ using 24 linear regressions to predict the 24 hourly Dutch electricity prices over the day using the price in the last hour of the previous day (and a constant) as a predictor.

- **Out-of-sample RMSE plot**



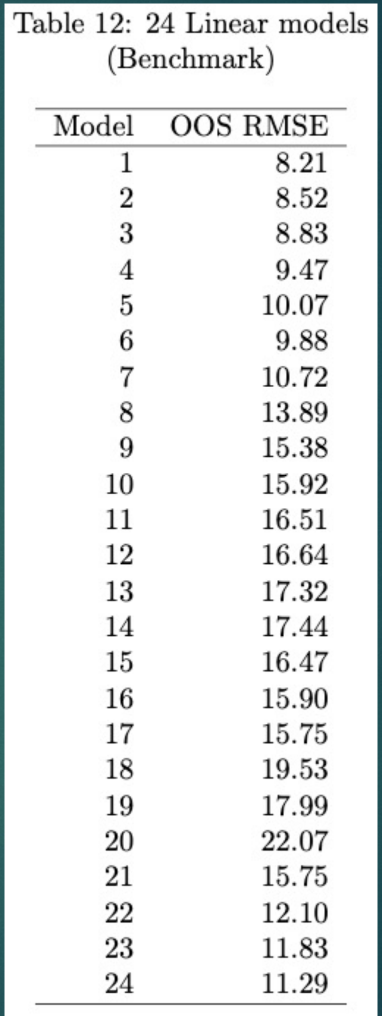


Table 12: 24 Linear models (Benchmark)

Model	OOS RMSE
1	8.21
2	8.52
3	8.83
4	9.47
5	10.07
6	9.88
7	10.72
8	13.89
9	15.38
10	15.92
11	16.51
12	16.64
13	17.32
14	17.44
15	16.47
16	15.90
17	15.75
18	19.53
19	17.99
20	22.07
21	15.75
22	12.10
23	11.83
24	11.29

- The heavy in-sample decrease in predictive power of the linear model due to the increase in volatility is also perfectly reflected in the out-of-sample RMSE plot. During the peak hours, the linear model tends to predict much worse which produces higher error rates, confirming its negative relationship with volatility.

Question 4.2 (1)

Machine learning models ~ Lasso model

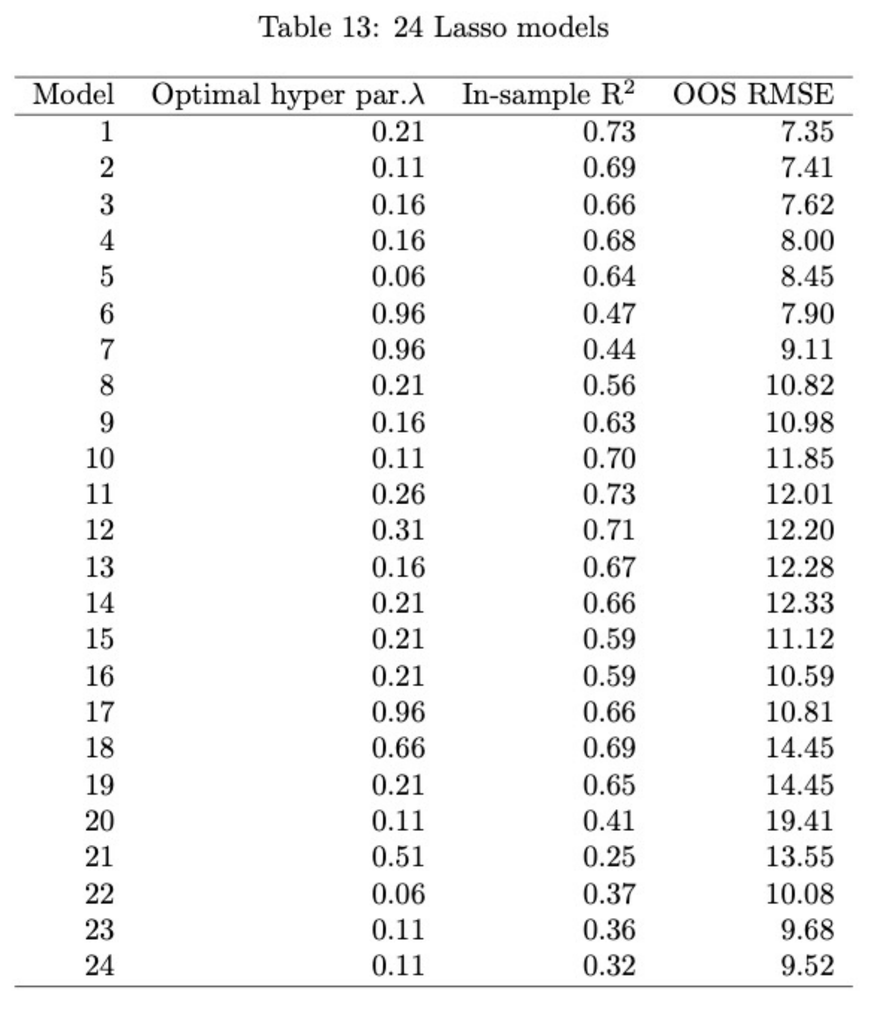
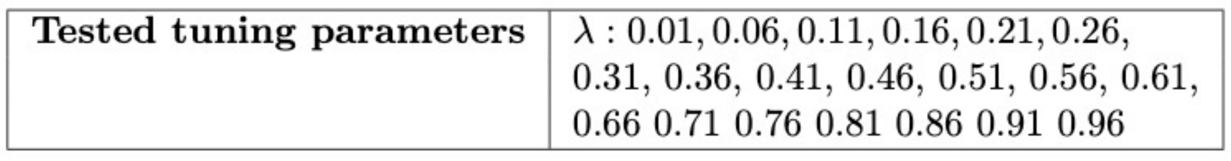


Table 13: 24 Lasso models

Model	Optimal hyper par.λ	In-sample R^2	OOS RMSE
1	0.21	0.73	7.35
2	0.11	0.69	7.41
3	0.16	0.66	7.62
4	0.16	0.68	8.00
5	0.06	0.64	8.45
6	0.96	0.47	7.90
7	0.96	0.44	9.11
8	0.21	0.56	10.82
9	0.16	0.63	10.98
10	0.11	0.70	11.85
11	0.26	0.73	12.01
12	0.31	0.71	12.20
13	0.16	0.67	12.28
14	0.21	0.66	12.33
15	0.21	0.59	11.12
16	0.21	0.59	10.59
17	0.96	0.66	10.81
18	0.66	0.69	14.45
19	0.21	0.65	14.45
20	0.11	0.41	19.41
21	0.51	0.25	13.55
22	0.06	0.37	10.08
23	0.11	0.36	9.68
24	0.11	0.32	9.52

- From the In-sample R-squared, it can be inferred that the further within the 24 hours we try to predict Dutch electricity prices, the less predictive power our lasso model seems to have which is reflected by the decrease in the R-squared.
- The out-of-sample RMSE increase after the 7th model can best be explained by the heavy increase in demand during the peak hours, which increases the volatility of the prices and hence also the error rates of our predictions.



Tested tuning parameters	$\lambda : 0.01, 0.06, 0.11, 0.16, 0.21, 0.26,$ 0.31, 0.36, 0.41, 0.46, 0.51, 0.56, 0.61, 0.66 0.71 0.76 0.81 0.86 0.91 0.96

Question 4.2 (2)

Machine learning models ~ Partial Least Squares

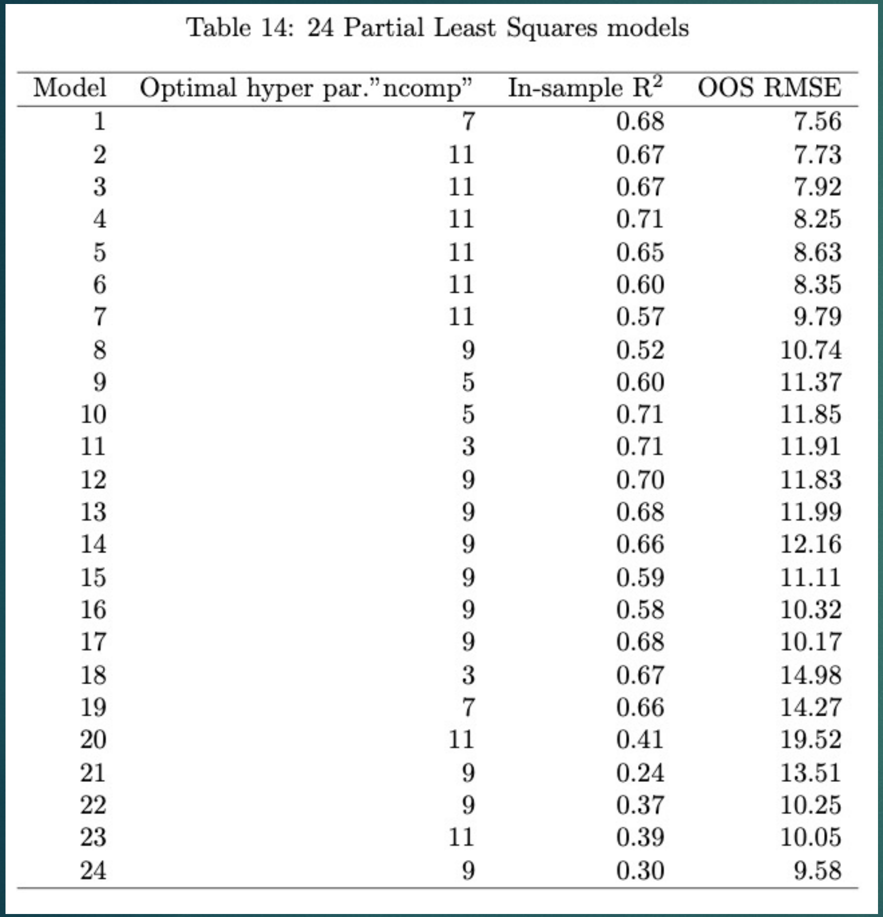
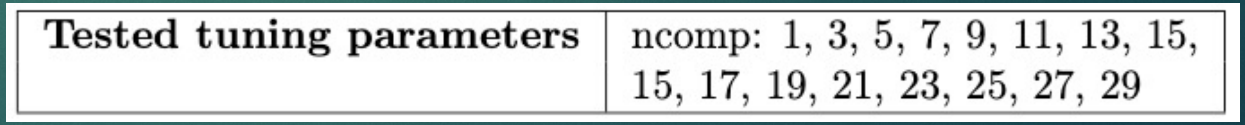


Table 14: 24 Partial Least Squares models

Model	Optimal hyper par."ncomp"	In-sample R^2	OOS RMSE
1	7	0.68	7.56
2	11	0.67	7.73
3	11	0.67	7.92
4	11	0.71	8.25
5	11	0.65	8.63
6	11	0.60	8.35
7	11	0.57	9.79
8	9	0.52	10.74
9	5	0.60	11.37
10	5	0.71	11.85
11	3	0.71	11.91
12	9	0.70	11.83
13	9	0.68	11.99
14	9	0.66	12.16
15	9	0.59	11.11
16	9	0.58	10.32
17	9	0.68	10.17
18	3	0.67	14.98
19	7	0.66	14.27
20	11	0.41	19.52
21	9	0.24	13.51
22	9	0.37	10.25
23	11	0.39	10.05
24	9	0.30	9.58

- Compared to the Lasso model, the In-sample R-squared stays for a long time stable until very late in the day for which then again can be inferred that the further away we try to predict Dutch electricity prices, the less predictive power our model seems to have.
- The out-of-sample RMSE seems to increase strongly when the peak hours start, in which heavy increase in demand increases the volatility and hence also increases the error rate of the predictions.



Tested tuning parameters	ncomp: 1, 3, 5, 7, 9, 11, 13, 15, 15, 17, 19, 21, 23, 25, 27, 29

Question 4.2 (3)

Machine learning models ~ Random Forest

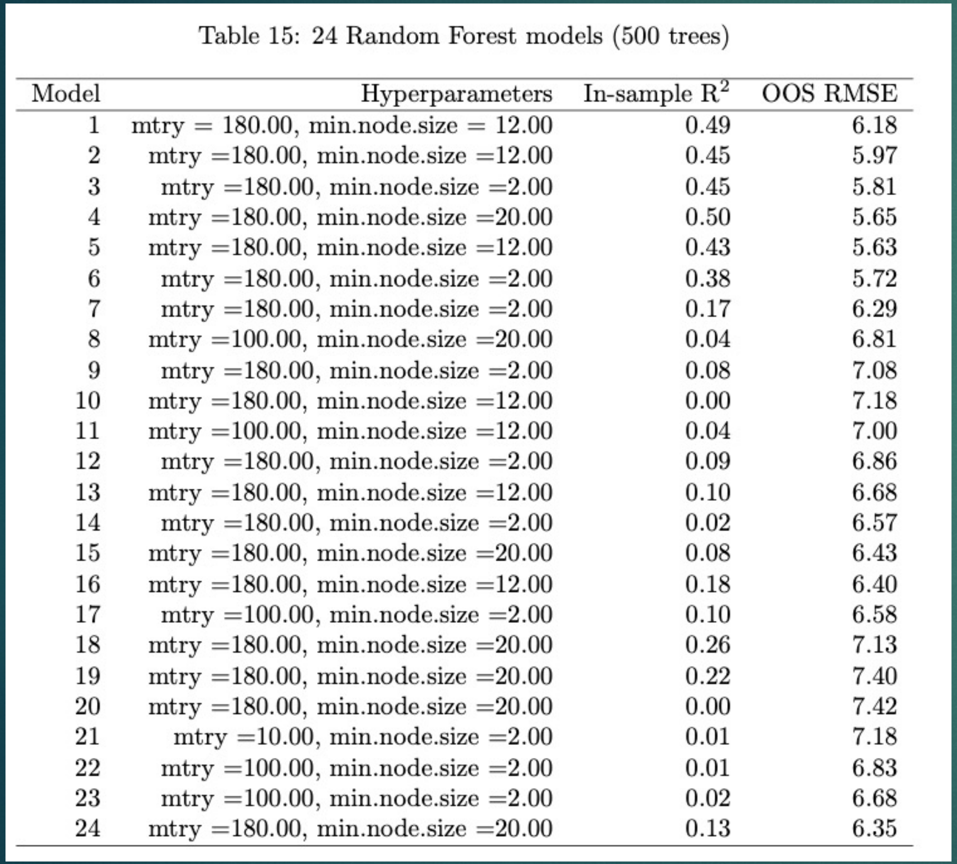
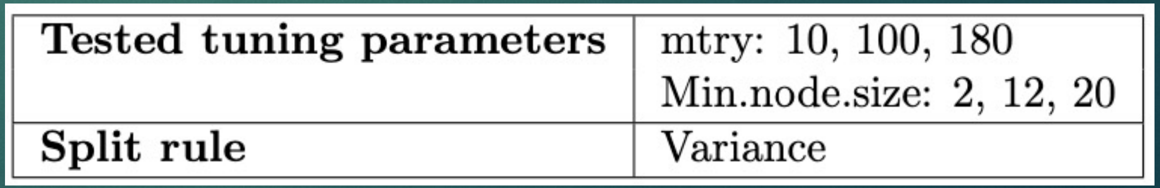


Table 15: 24 Random Forest models (500 trees)

Model	Hyperparameters	In-sample R^2	OOS RMSE
1	mtry = 180.00, min.node.size = 12.00	0.49	6.18
2	mtry =180.00, min.node.size =12.00	0.45	5.97
3	mtry =180.00, min.node.size =2.00	0.45	5.81
4	mtry =180.00, min.node.size =20.00	0.50	5.65
5	mtry =180.00, min.node.size =12.00	0.43	5.63
6	mtry =180.00, min.node.size =2.00	0.38	5.72
7	mtry =180.00, min.node.size =2.00	0.17	6.29
8	mtry =100.00, min.node.size =20.00	0.04	6.81
9	mtry =180.00, min.node.size =2.00	0.08	7.08
10	mtry =180.00, min.node.size =12.00	0.00	7.18
11	mtry =100.00, min.node.size =12.00	0.04	7.00
12	mtry =180.00, min.node.size =2.00	0.09	6.86
13	mtry =180.00, min.node.size =12.00	0.10	6.68
14	mtry =180.00, min.node.size =2.00	0.02	6.57
15	mtry =180.00, min.node.size =20.00	0.08	6.43
16	mtry =180.00, min.node.size =12.00	0.18	6.40
17	mtry =100.00, min.node.size =2.00	0.10	6.58
18	mtry =180.00, min.node.size =20.00	0.26	7.13
19	mtry =180.00, min.node.size =20.00	0.22	7.40
20	mtry =180.00, min.node.size =20.00	0.00	7.42
21	mtry =10.00, min.node.size =2.00	0.01	7.18
22	mtry =100.00, min.node.size =2.00	0.01	6.83
23	mtry =100.00, min.node.size =2.00	0.02	6.68
24	mtry =180.00, min.node.size =20.00	0.13	6.35

- When approaching the peak hours, the in-sample R-squared seems to decrease rapidly. Suggesting that the Random Forest model fails (almost) completely to capture the variation in the training data during these hours. This could be explained by the increase in volatility.
- Surprisingly, the out-of-sample RMSE stays rather stable and lower compared to the other models. This could suggest that the RF model has made perfectly the bias-variance trade-off, meaning it avoided overfitting which results in better predictions in the test set.

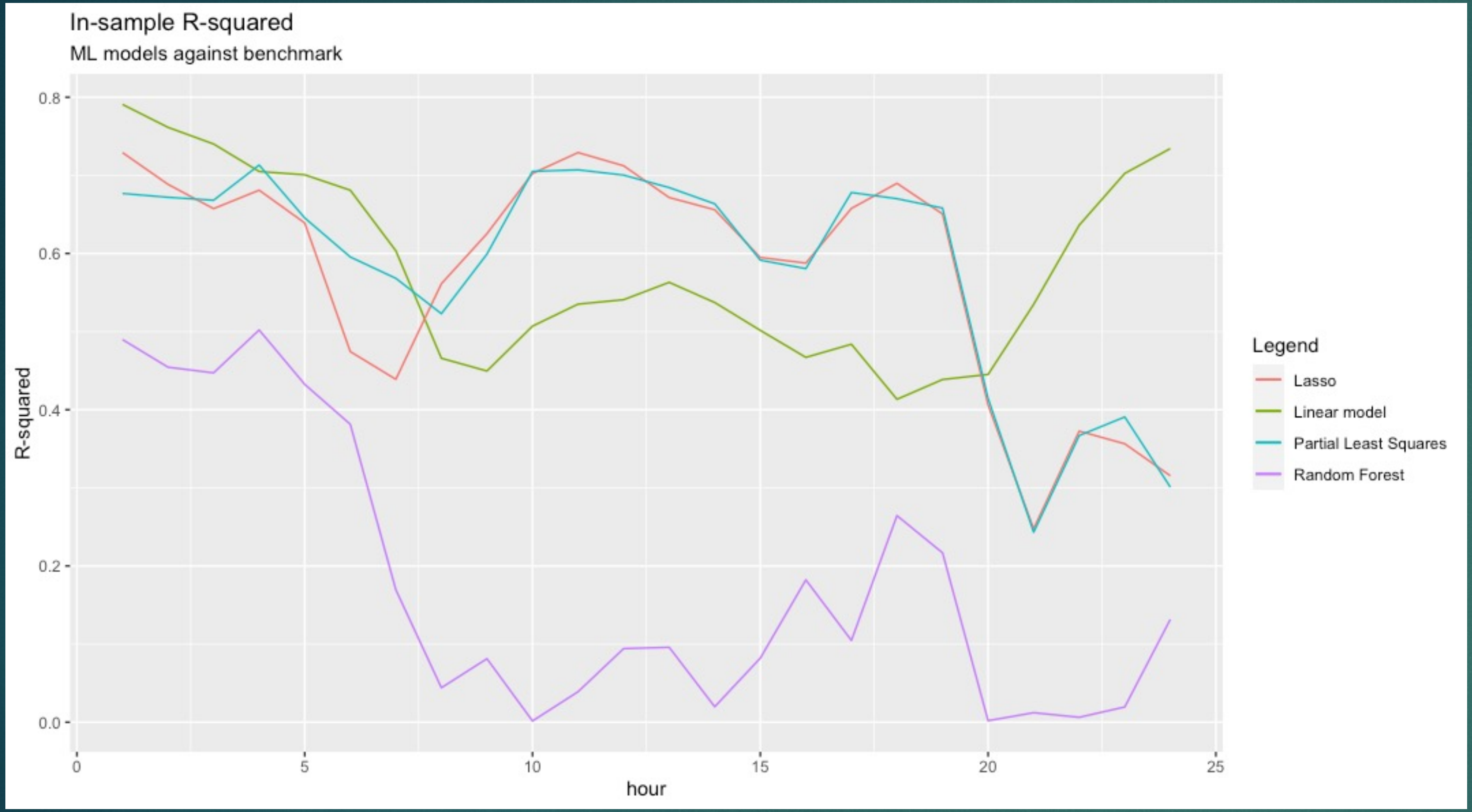


Tested tuning parameters	mtry: 10, 100, 180 Min.node.size: 2, 12, 20
Split rule	Variance

Question 4.2 (4)

Machine learning models

- **In-sample R-squared plot**

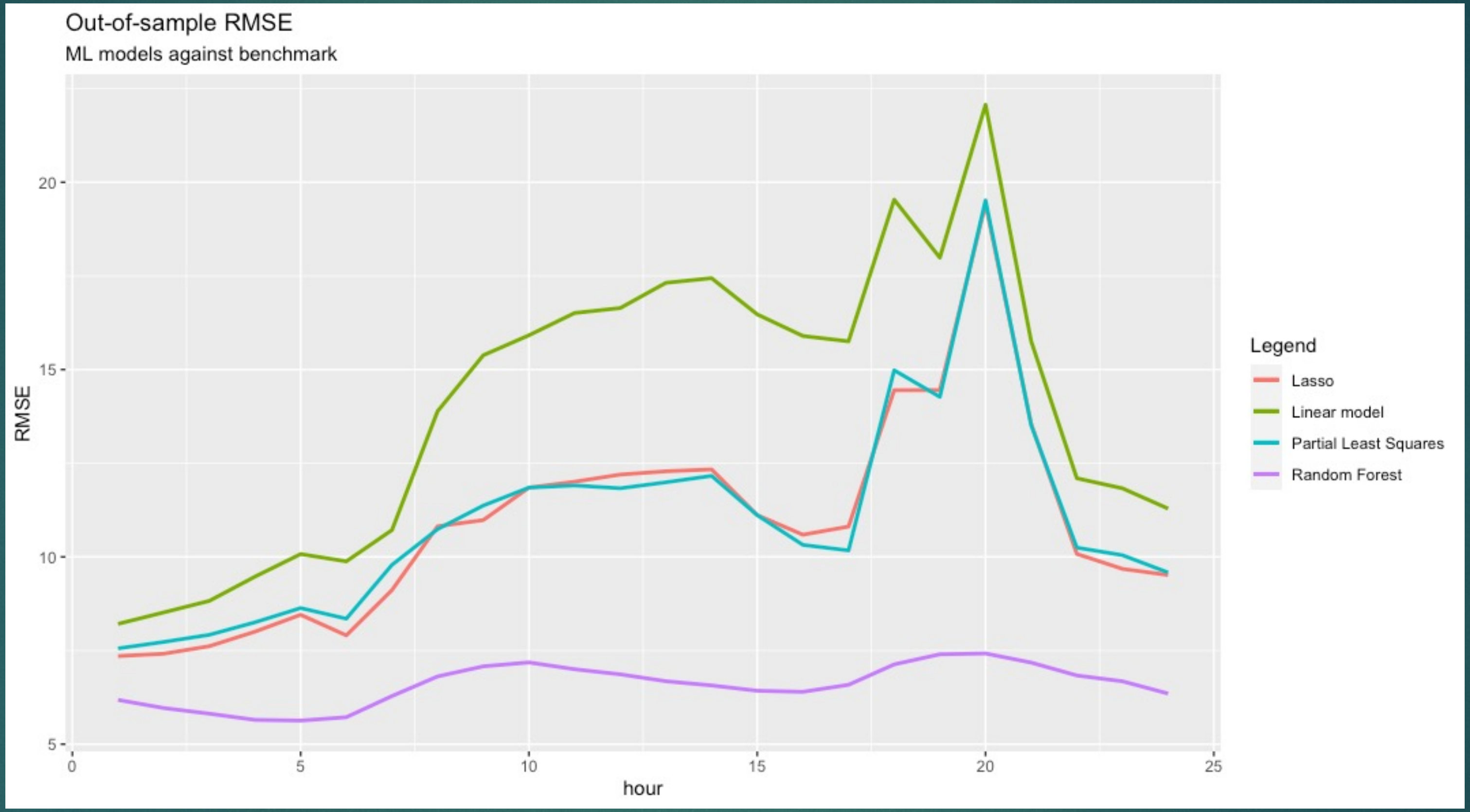


- The benchmark (linear model), as well as the lasso and PLS model follow a similar pattern. Not surprisingly, since these models are regression models that aim to minimize the squared errors, each in its own way.
- The Random Forest model seems to fit very bad the test data, the pattern is very different from the other models since this model is not a regression model, but rather a tree based model which is known to explain the variation better for classification problems.

Question 4.2 (5)

Machine learning models

- **Out-of-sample RMSE plot**



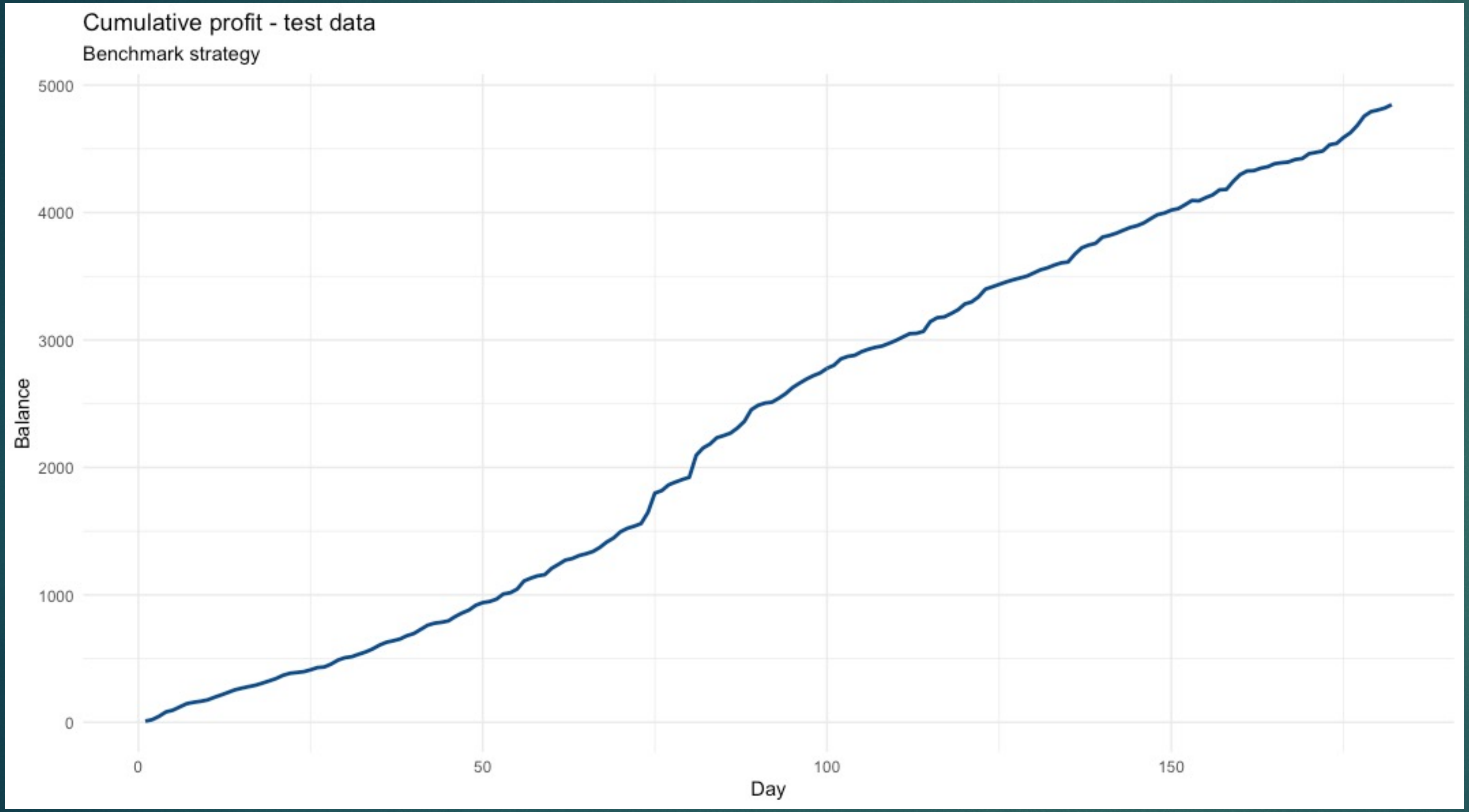
- Despite the low in-sample R-squared of the random forest, it outperforms the regression based models in the test set. The random forest model did a very good job in finding a good “bias-variance” trade-off, leading to more accurate out-of-sample predictions.

Question 5: Trading strategy

Question 5.1

Benchmark strategy

- **Cumulative profit on test data plot**

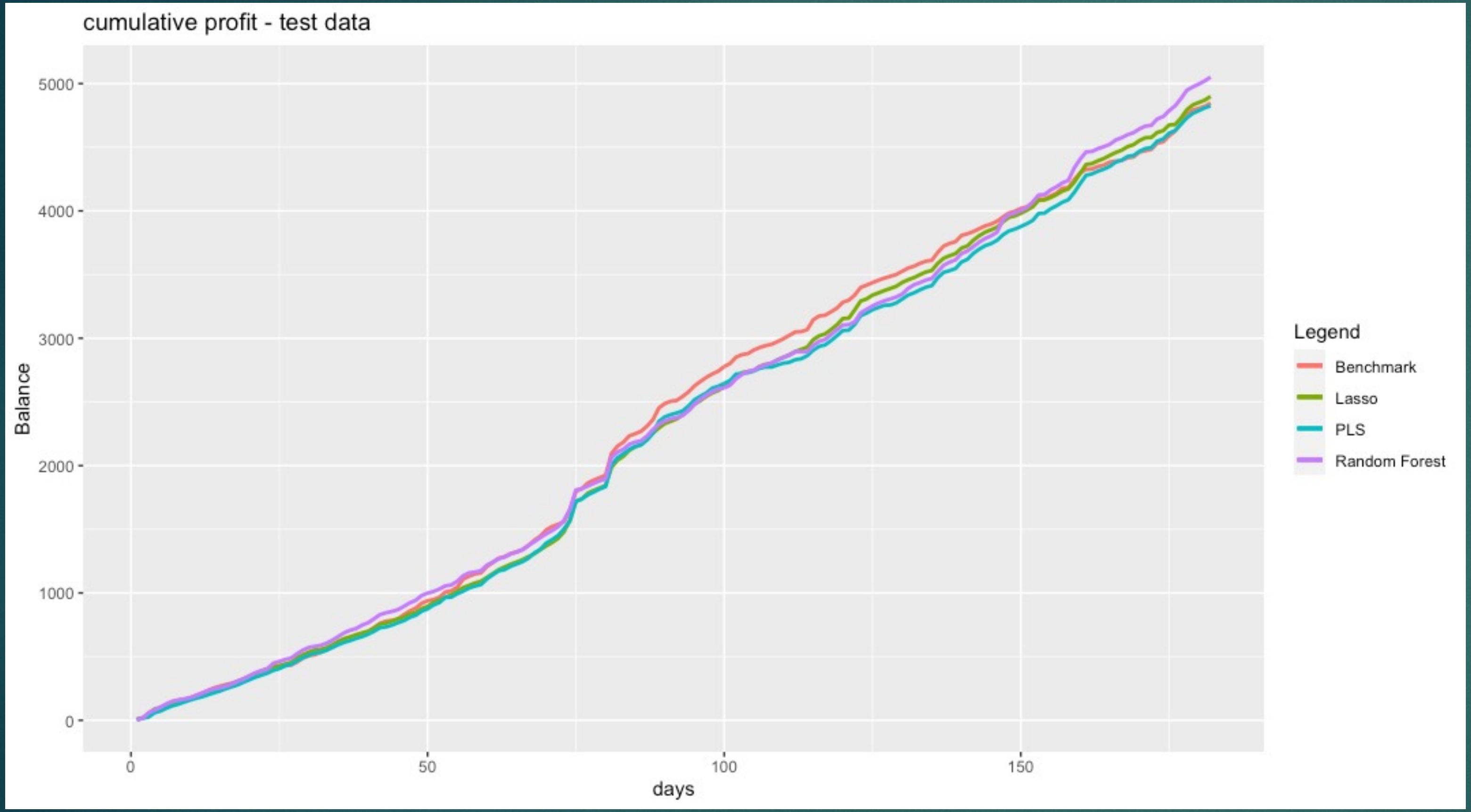


- Final balance : 4846.76
- Daily volatility: 11.76 %

Question 5.2

Machine learning models

- **Cumulative profit on test data plot**

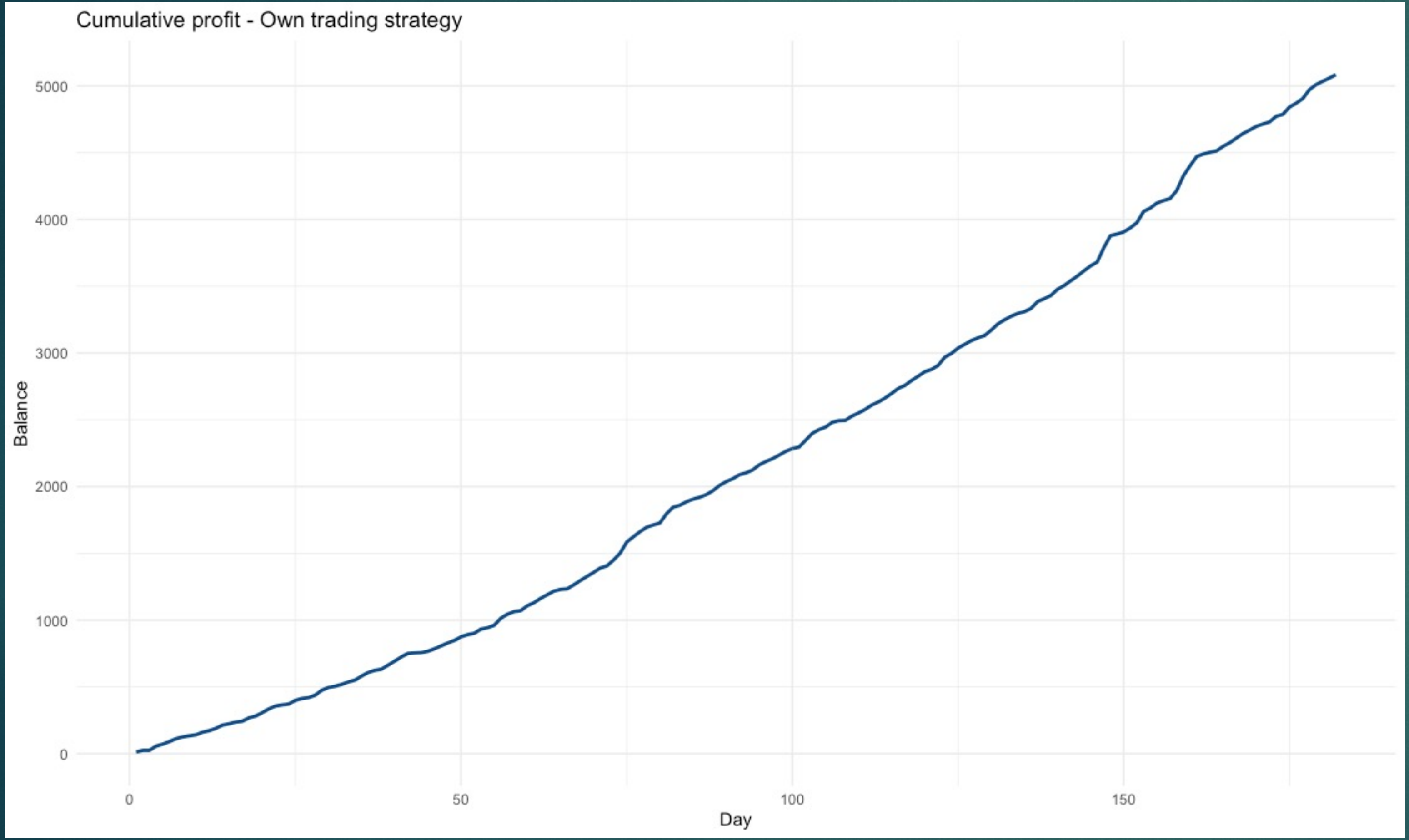


- Lasso
  - Final balance : 4899.06
  - Daily volatility: 10.59 %
- Partial Least Squares
  - Final balance : 4825.03
  - Daily volatility: 11.35 %
- Random Forest
  - Final balance : 5050.45
  - Daily volatility: 11.63 %
  - **Best model**

Question 5.3 (1)

Build your own trading strategy

- **Cumulative profit on test data plot**



- Results
  - Final balance : 5084.24
  - Daily volatility: 9.70 %
  - Outperforms all strategies of previous question

- Approach
  - See next slide

Question 5.3 (2)

Build your own trading strategy

- Brief explanation of approach
  - Step 1: Deciding when and how many times a day to charge and discharge battery
    - How many times → two charge and discharge cycles:
      - 1) Charge at 6 and discharge at 10
      - 2) Charge at 13 and discharge at 18
      - Motivation: Based on the OOS predictions of my best model (random forest), these are the cycles in which I can maximize profit. This strategy also makes sure that after charging, I discharge first before charging again.
  - Step 2: Computing cumulative profit
    - Final balance : 5084.24
    - Daily volatility: 9.70 %