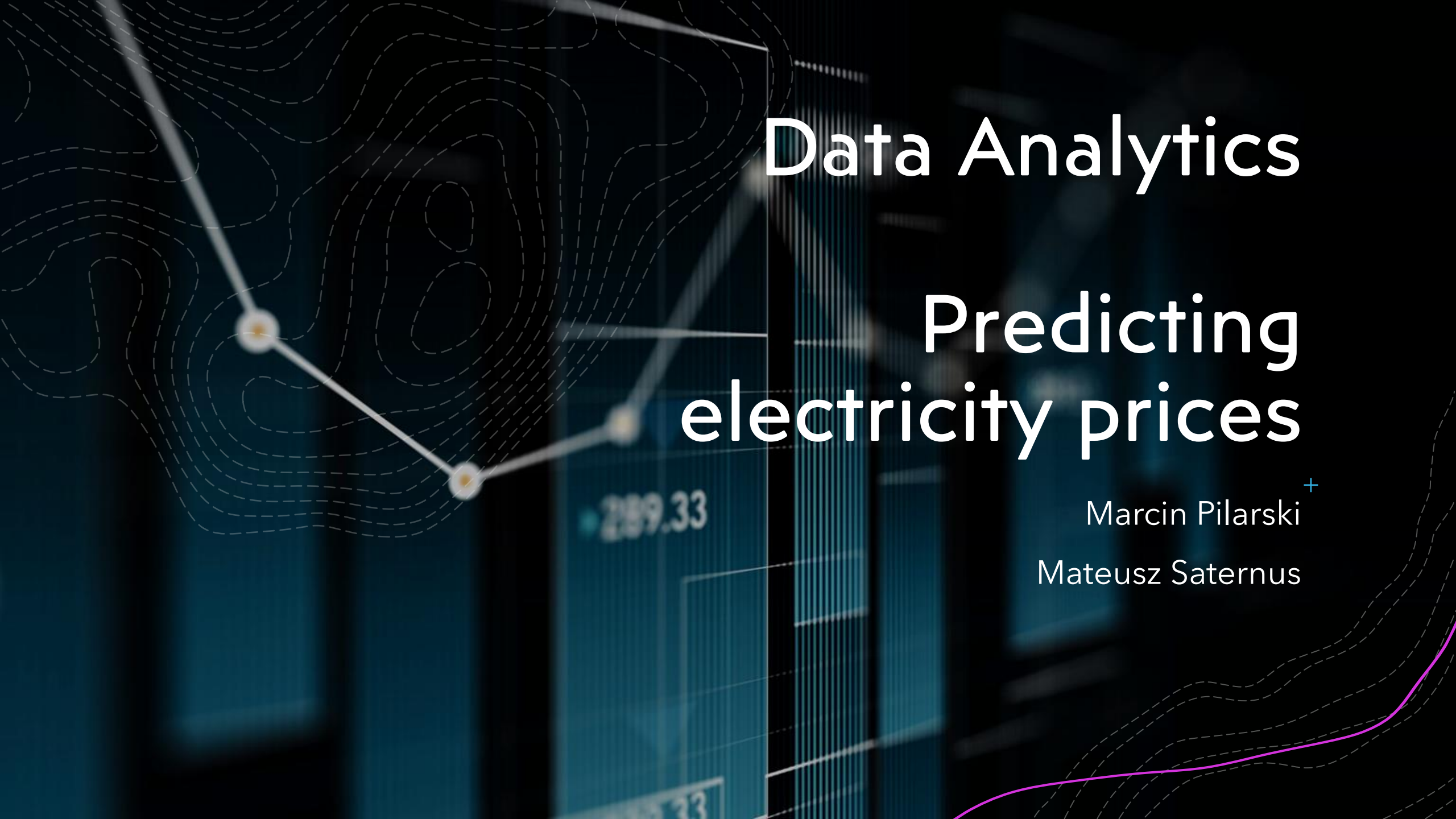


Data Analytics

Predicting electricity prices

Marcin Pilarski

Mateusz Saternus

Problem formulation

- The energy market plays a crucial role in the economic development and sustainability of nations worldwide. The energy market in Poland faces multifaceted challenges, including fluctuating energy prices, increasing demand, changing regulatory frameworks, and the need to transition towards a more sustainable energy mix. These complexities pose significant difficulties for market participants, policymakers, and investors in making informed decisions regarding energy production, distribution, and consumption. To tackle these challenges effectively, it is crucial to understand the underlying patterns and interdependencies within the market, which can be achieved through statistical modelling.

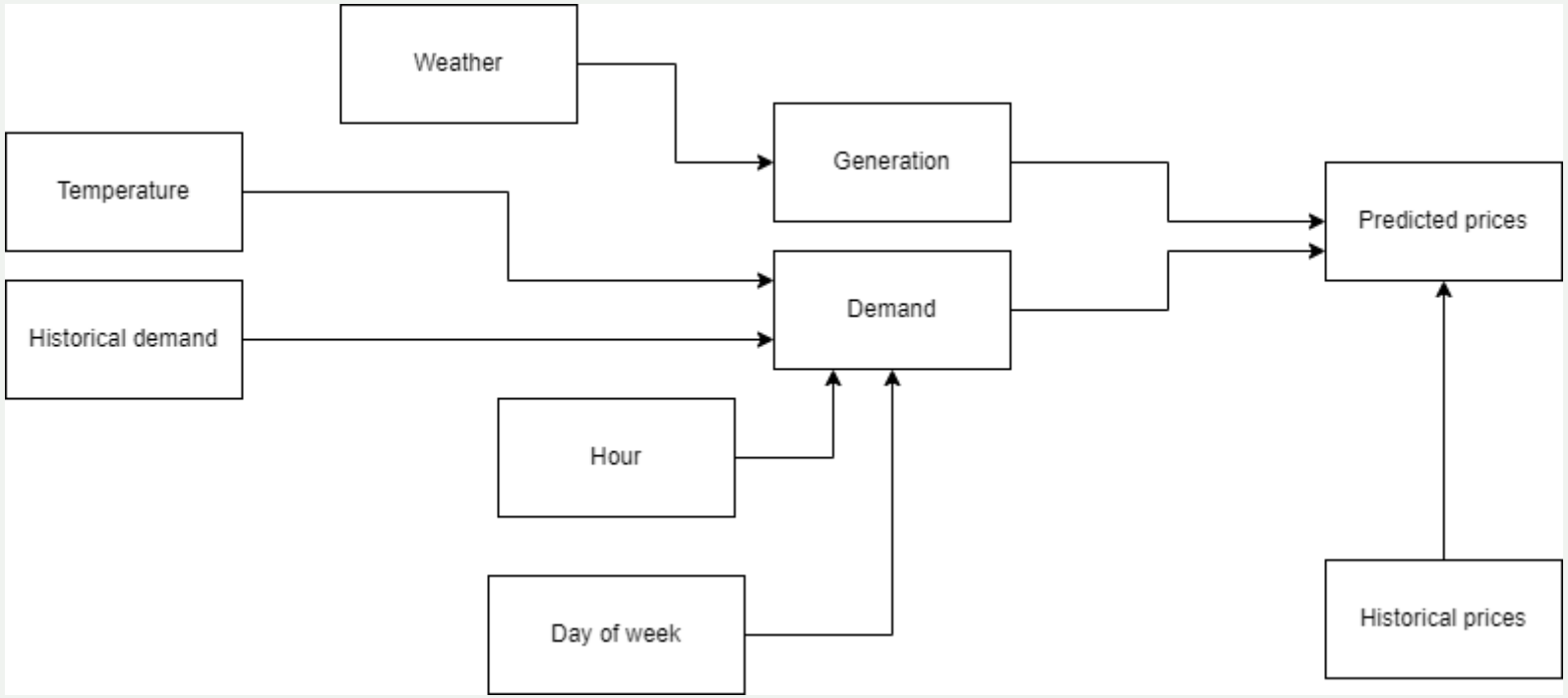
Problem formulation – potencial use cases

- The statistical modelling of the energy market in Poland offers a wide range of potential use cases across different sectors. For energy companies, the model can assist in optimizing production levels, pricing strategies, and investment decisions by providing insights into demand patterns and market conditions. Regulators and policymakers can leverage the model to evaluate the effectiveness of existing policies, design new regulations, and identify areas for promoting renewable energy sources and energy efficiency. Investors can use the model to assess the financial viability and risks associated with energy projects, aiding in informed decision-making.

Problem formulation – data sources

- Data used in model comes from official databases. First database used in project is PVgis, which is tool released by European Commission. (https://re.jrc.ec.europa.eu/pvg_tools/en/) It helps to plan investments in solar and wind energy sources. The data contains irradiance, wind speed, temperature.
- The information about energy market such as energy prices and system load come from polish energy system operator (PSE). (https://www.pse.pl/obszary-dzialalnosci/rynek-energii/ceny-i-ilosc-energii-na-rynku-bilansujacym; https://www.pse.pl/dane-systemowe/funkcjonowanie-kse/raporty-dobowe-z-pracy-kse/wielkosci-podstawowe)

Problem formulation- Directed Acyclic Grap



Problem formulation - confoundings

- There are few confoundings detected in the model.
- The pipe type is on relation of weather->generation->price.
- There are also fork type confoundings:
  - Hour -> Load <- Day of week
  - Generation -> Price <- Historical price

Data preprocessing

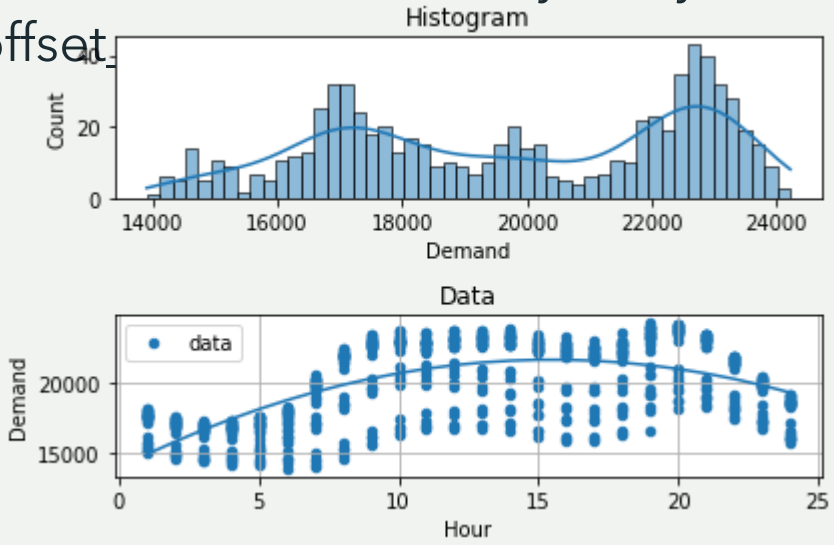
- The data about the weather was downloaded from PVgis database. Although it contains all the weather values that were useful for us, it allows only to collect data locally and we need information for the whole country. Because of that we have chosen 22 points in Poland and calculated mean value at the time. Points are shown on the map. Data is collected every hour and was downloaded separately for year 2019 and 2020.
- The data about the energy market was downloaded from PSE polish energy system operator. It was downloaded for march 2019 and march 2020 separately. It contained information about the system load and balancing market energy prices. The data was slightly changed to allow easier operations on dates.
- Data from 2019 was used for prior and from 2020 for posterior

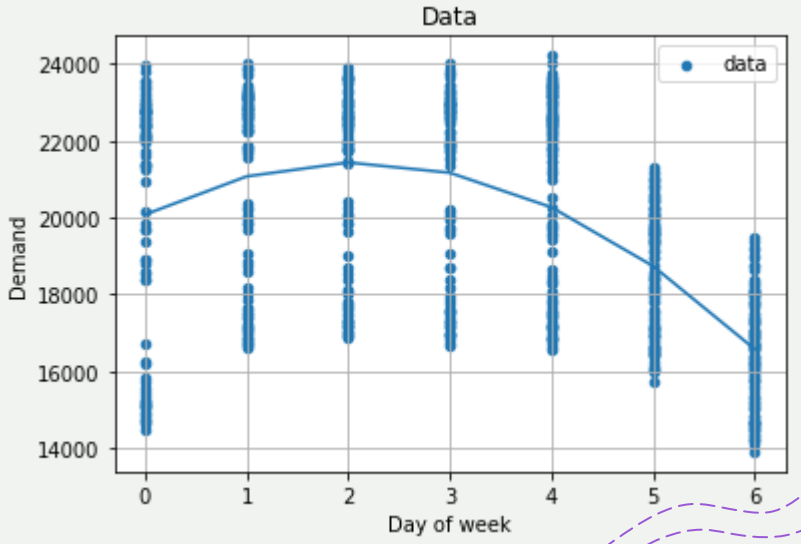
Data preprocessing

Collected data was loaded into DataFrames to allow easy manipulation of the columns. Most of them didn't require any modification, so the most important change that had to be made, was ensuring the same format of the columns corresponding to time. To that end, every DataFrame's time column was modified to be a datetime object. Additionally, a column Day of the week was added to the DataFrame with the demand data, using the time information in a clever way. The data from 2020 about irradiance also had to be summed up, as it was divided into three parts. The last modification was limiting the data to one month.

Model 1 – Demand

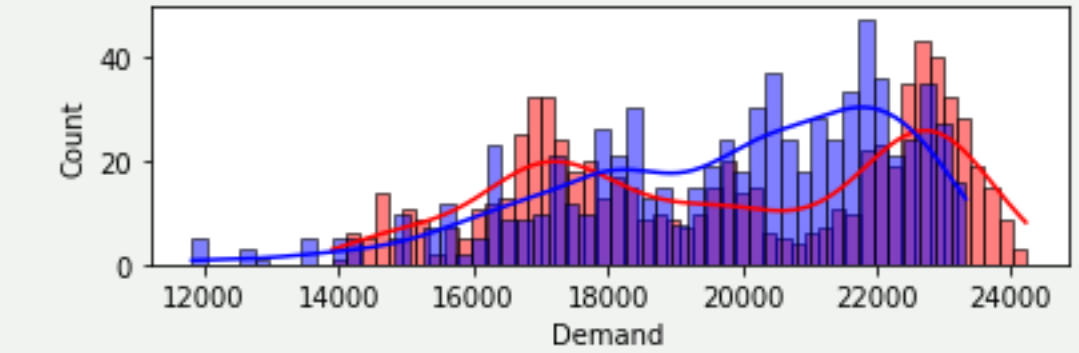
- The load has been fitted with 2nd degree polynomial for each dependence on hour and day of week
- Formula: demand = a_day * day ^2 + b_day * day + a_hour * hour ^ 2 + b_hour * hour + offset

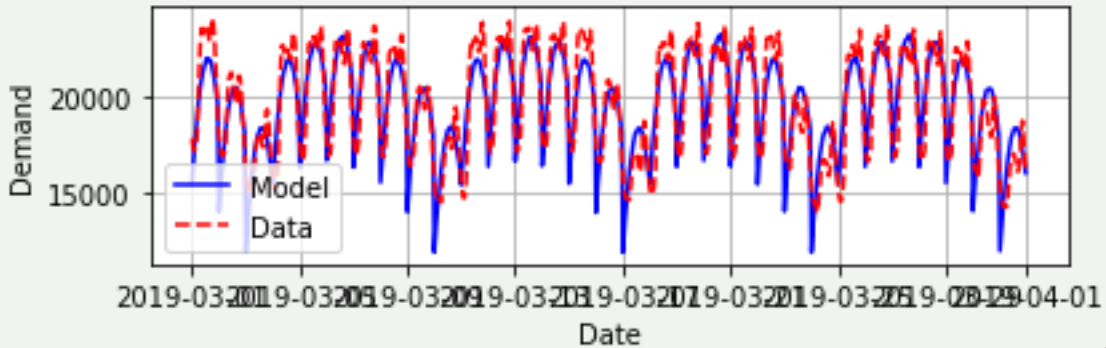




Model 1 – Demand - Prior

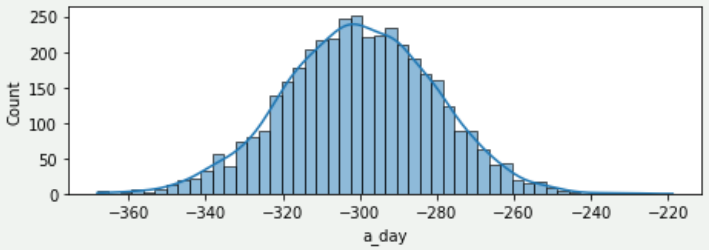
- Fitting time data with 2nd degree polynomial allowed us to get samples with distribution shown below
- Value of error was:
  RMSE = 1352.998 / (max possible) 19982.844

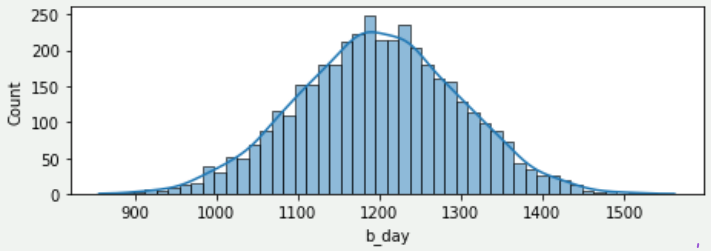


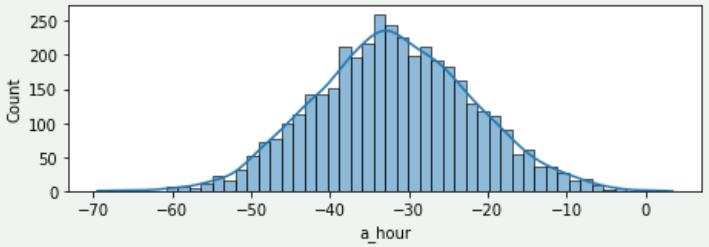


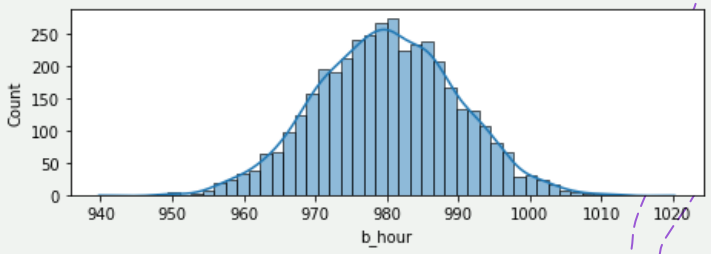
Model 1 – Demand – Prior Summary

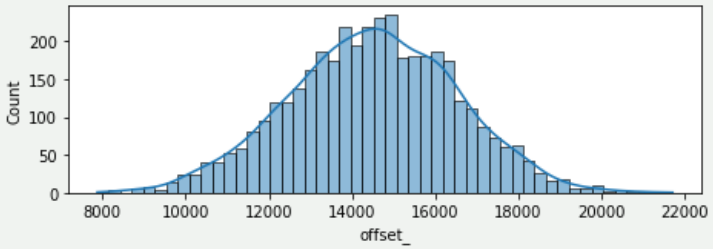
name	Mean	MCSE	StdDev	5%	50%	95%	N_Eff	N_Eff/s	R_hat
lp__	0.0	NaN	0.0	0.0	0.0	0.0	NaN	NaN	NaN
a_hour	-32.0	0.16	10.0	-49.0	-32.0	-15.0	4200.0	650.0	1.0
b_hour	980.0	0.15	9.9	960.0	980.0	1000.0	4200.0	650.0	1.0
offset_	14000.0	32.00	2000.0	11000.0	15000.0	18000.0	4000.0	620.0	1.0
a_day	-300.0	0.32	20.0	-330.0	-300.0	-270.0	3700.0	570.0	1.0
...	...	...	...	...	...	...	...	...	...
demand[739]	17680.0	72.00	4590.0	10173.0	17570.0	25313.0	4012.0	621.0	1.0
demand[740]	17344.0	79.00	4886.0	9343.0	17336.0	25360.0	3817.0	590.0	1.0
demand[741]	16908.0	85.00	5336.0	8117.0	16830.0	25644.0	3907.0	604.0	1.0
demand[742]	16492.0	92.00	5773.0	6734.0	16566.0	25916.0	3916.0	606.0	1.0
demand[743]	16013.0	108.00	6317.0	5847.0	15916.0	26478.0	3392.0	525.0	1.0





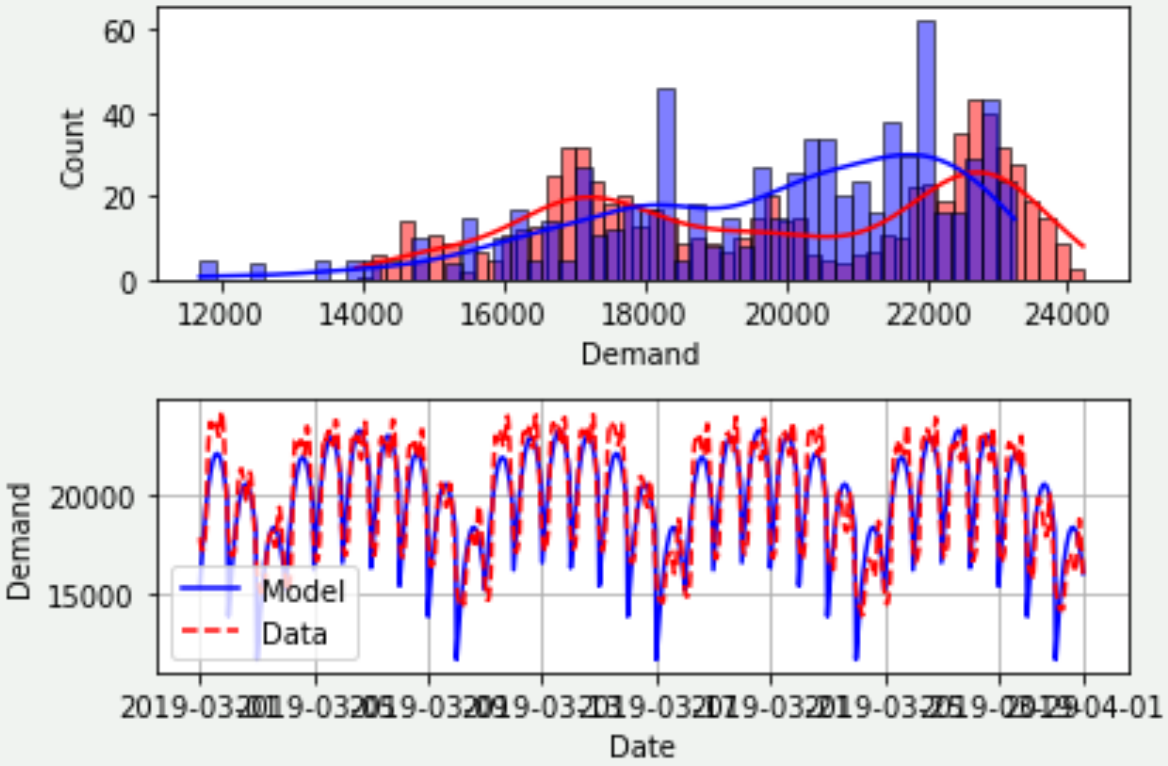






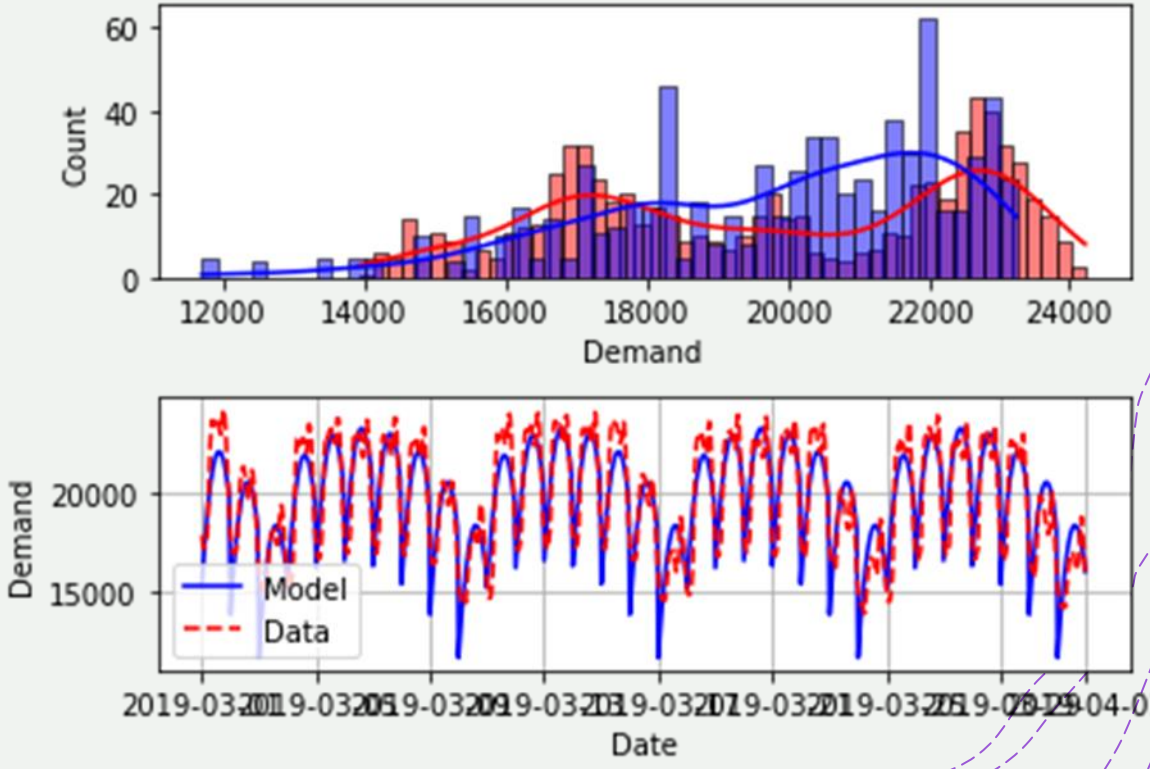
Model 1 – Demand – Posterior

- Posterior modeling gave the results shown on graph
- Value of error was: RMSE = 1348.9/ (max possible) 19982.8



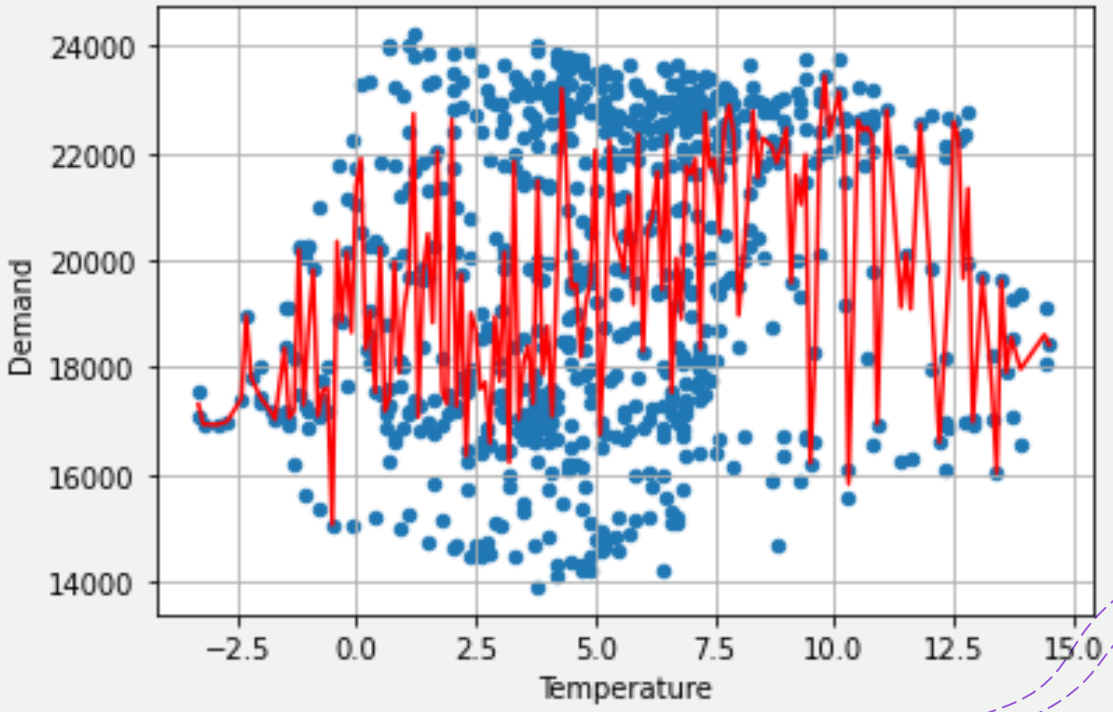
Model 1 – Demand – Posterior analysis

- Posterior predictive samples are quite consistent with the data, but there are more of them in middle range (18k – 22k)
- Also few samples appears in low range (below 14k), which does not exist in real data



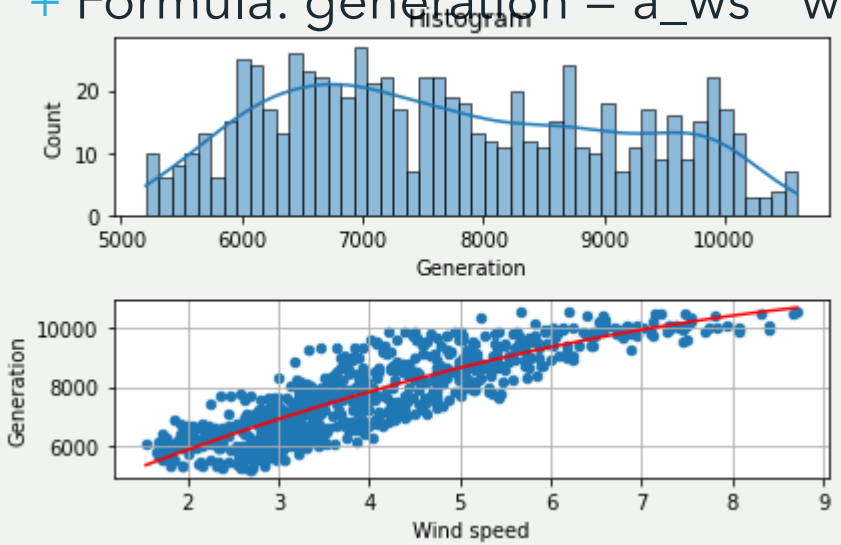
Temperature’s influence on the demand

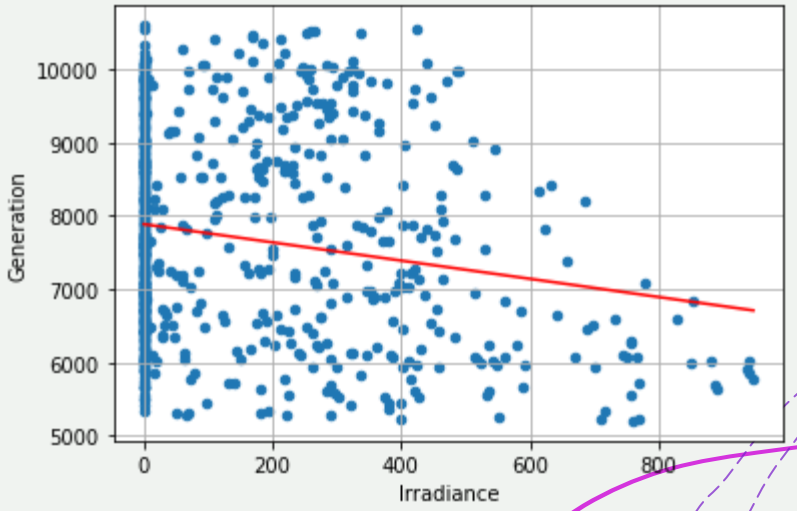
- It seems that there is no influence of temperature on the demand, contrary to the project's assumptions.
- Therefore, the influence of temperature was omitted in models



Model 1 - Generation

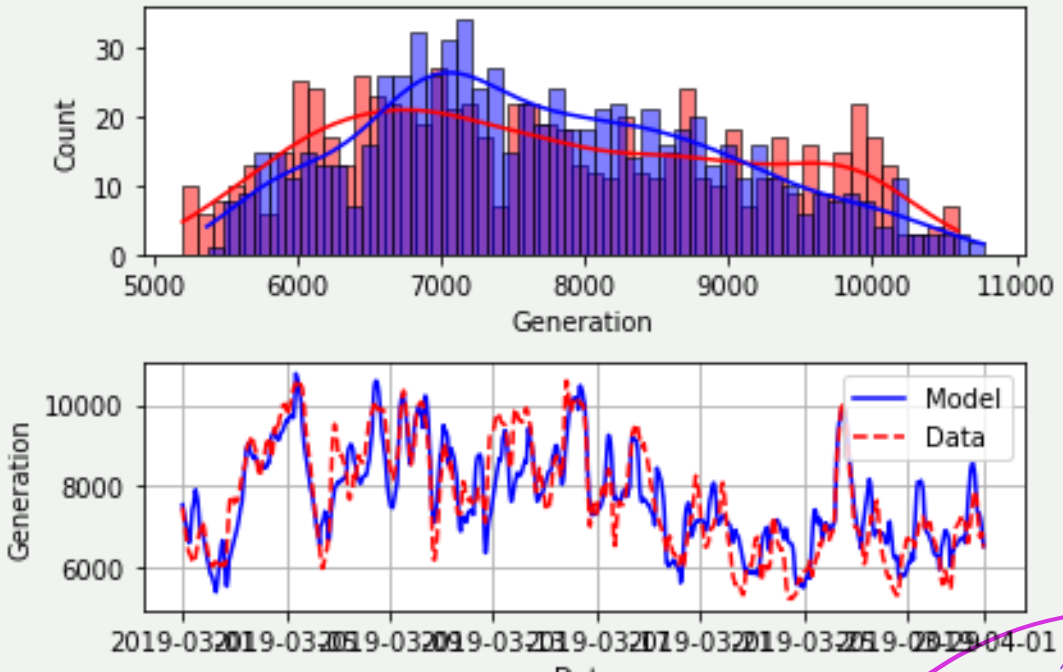
- Generation dependence on wind and solar irradiance has been tested
- As shown below, there is strong dependence on wind speed and almost no dependance on irradiance
- Wind speed has been fitted with quadratic function and irradiance with linear
- Therefore we omitted irradiance influence on generation in 1st model – it is added in the 2nd
- Formula: generation = a_ws * wind + b_ws * wind^2 + offset_





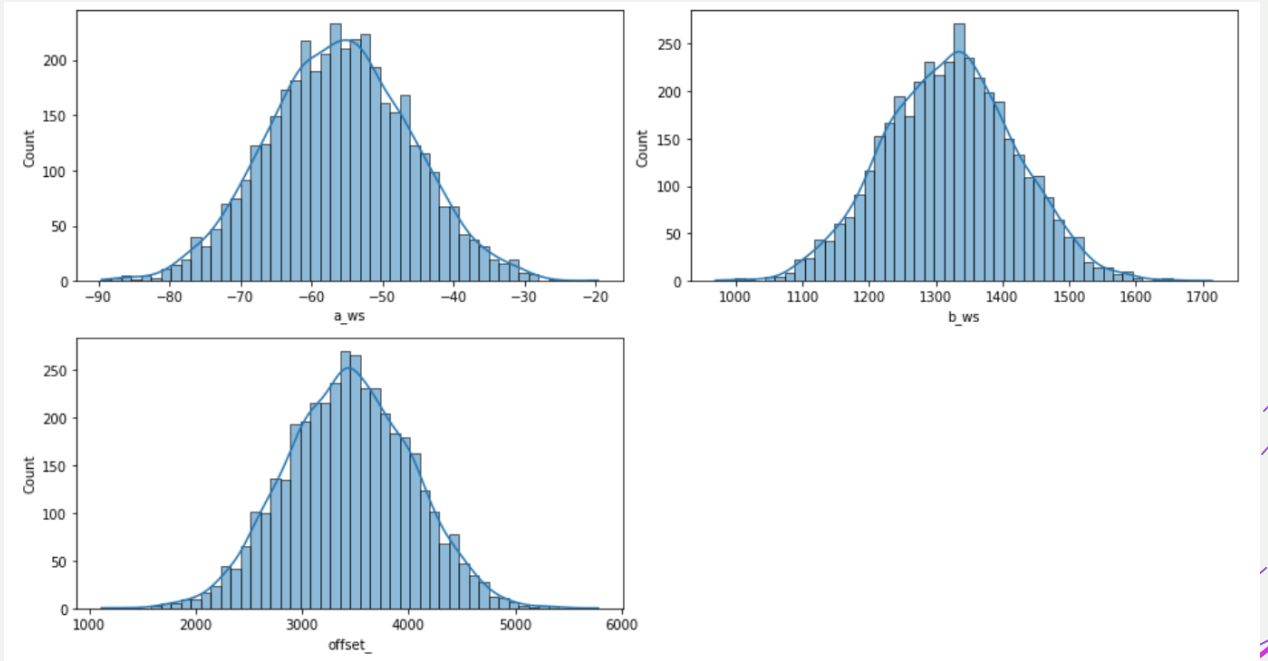
Model 1 – Generation - Prior

- The prior distribution for Generation in 1st model is shown below
- The value of error was: RMSE = 694.38 / (max possible) 7836.45



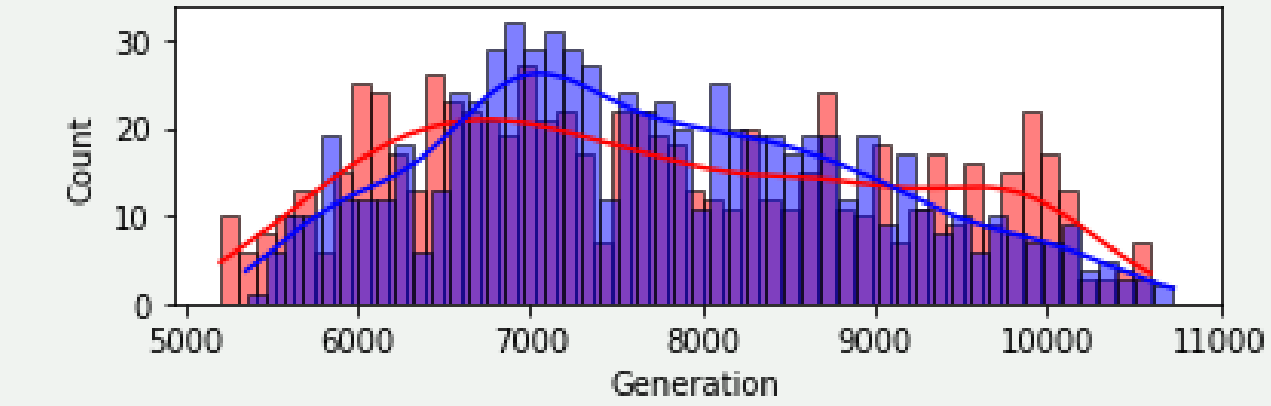
Model 1 – Generation – Prior summary

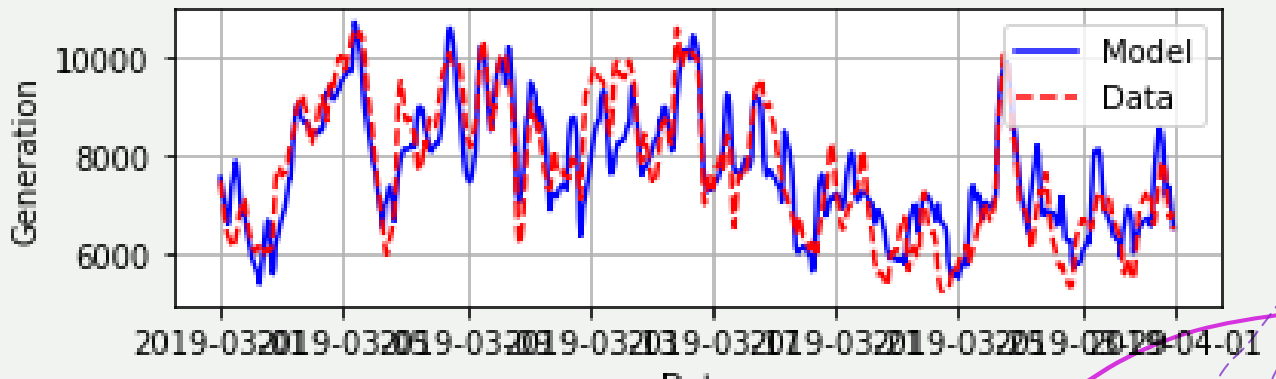
name	Mean	MCSE	StdDev	5%	50%	95%	N_Eff	N_Eff/s	R_hat
lp__	0.0	NaN	0.0	0.0	0.0	0.0	NaN	NaN	NaN
a_ws	-56.0	0.16	10.0	-72.0	-56.0	-40.0	3700.0	800.0	1.0
b_ws	1300.0	1.60	99.0	1200.0	1300.0	1500.0	3900.0	840.0	1.0
offset_	3400.0	9.20	590.0	2500.0	3400.0	4400.0	4100.0	890.0	1.0
generation[1]	7534.0	19.00	1248.0	5494.0	7503.0	9633.0	4219.0	910.0	1.0
...	...	...	...	...	...	...	...	...	...
generation[739]	7316.0	19.00	1225.0	5360.0	7294.0	9358.0	4107.0	886.0	1.0
generation[740]	7237.0	20.00	1216.0	5239.0	7257.0	9240.0	3854.0	832.0	1.0
generation[741]	7074.0	19.00	1225.0	5034.0	7093.0	9083.0	4003.0	864.0	1.0
generation[742]	6770.0	19.00	1193.0	4828.0	6756.0	8751.0	3888.0	839.0	1.0
generation[743]	6511.0	20.00	1206.0	4505.0	6496.0	8493.0	3736.0	806.0	1.0



Model 1 – Generation – Posterior

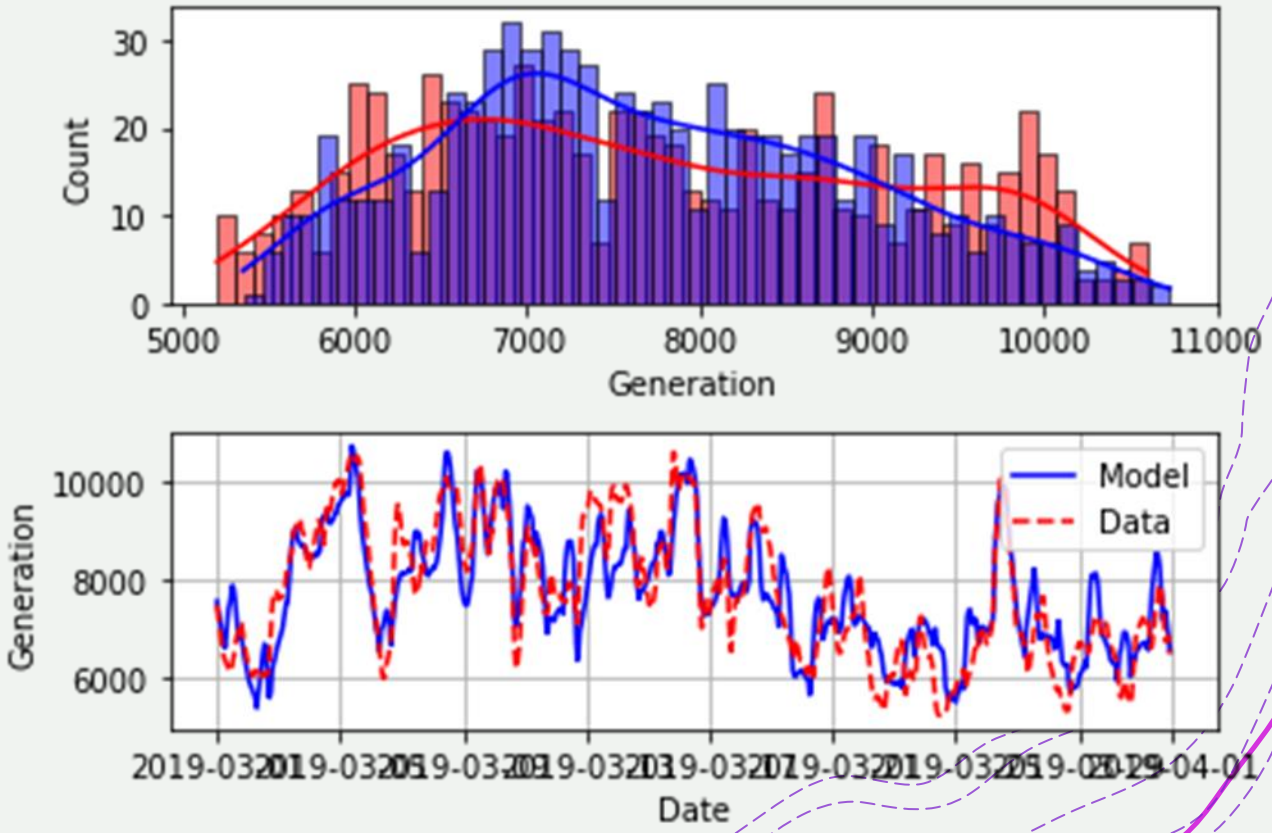
- Posterior modeling gave the results shown on graph
- The error value was: RMSE = 694.48 / (max possible) 7836.45





Model 1 – Generation – Posterior analysis

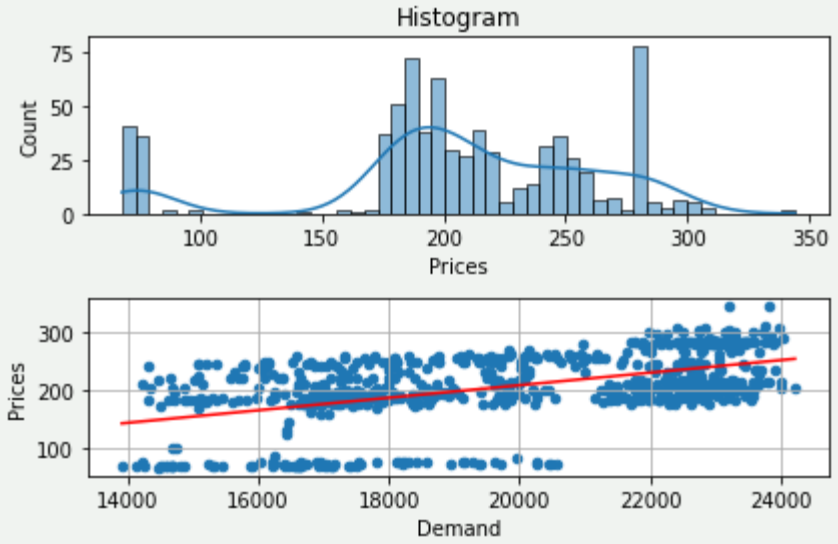
- The posterior predictive distribution is suprisingly good consistent with the data
- There are a bit more samples in middle range and fewer in distribution ends
- However values of samples do not exceed range of real data

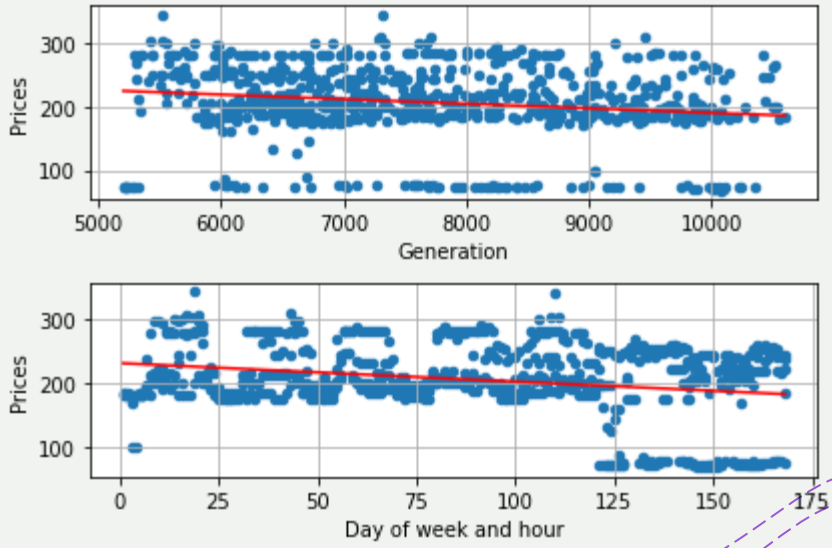


Model 1 - Prices

+ We determined the relations with load, generation and day of week with linear functions.

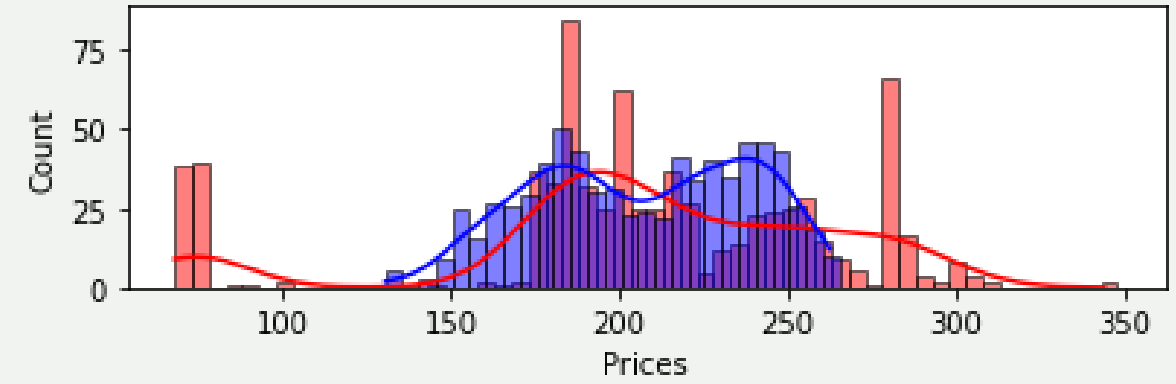
+ Formula: prices = a_demand * demand + a_generation * generation + offset_

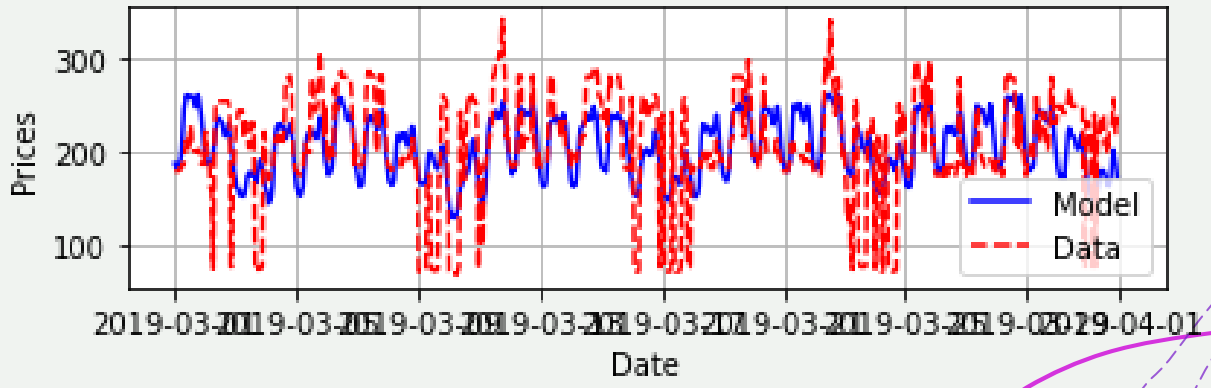




Model 1 – Prices - Prior

- Modelling Prior for prices in 1st Model gave results shown below
- The error value was: RMSE = 48.18 / (max possible) 214.47

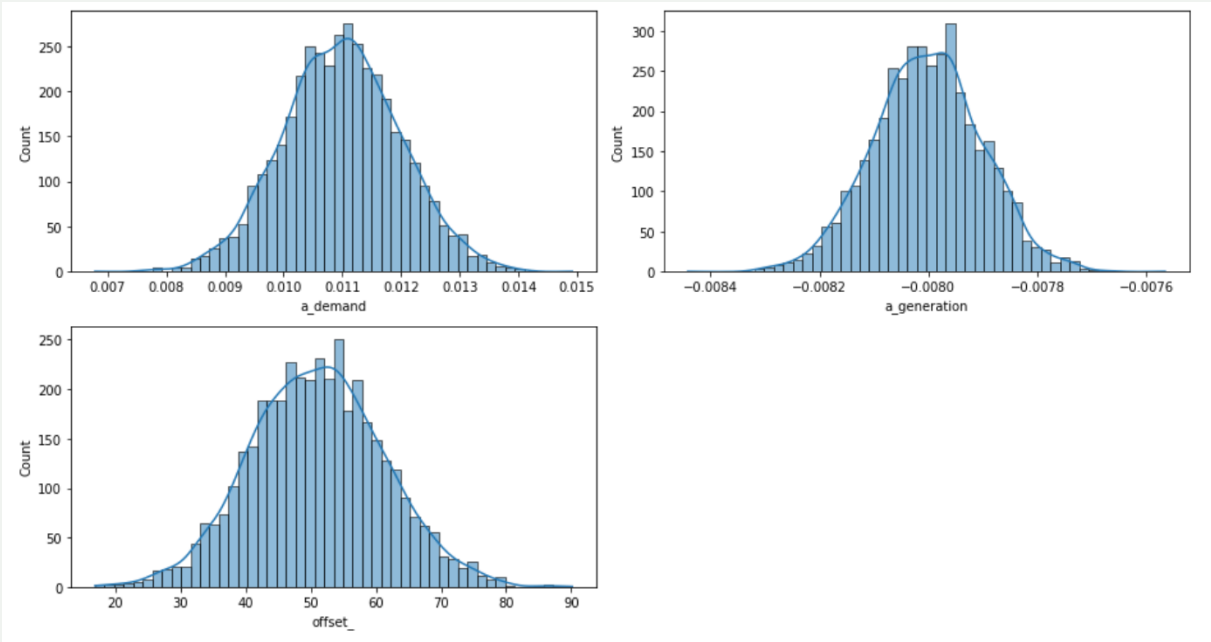




Model 1 – Prices – Prior summary

name	Mean	MCSE	StdDev	5%	50%	95%	N_Eff	\
lp__	0.000	NaN	0.0000	0.0000	0.000	0.0000	NaN	
a_demand	0.011	0.000016	0.0010	0.0094	0.011	0.0130	4100.0	
a_generation	-0.008	0.000002	0.0001	-0.0082	-0.008	-0.0078	3800.0	
offset_	51.000	0.160000	10.0000	34.0000	51.000	68.0000	4000.0	
prices[1]	187.000	1.600000	103.0000	19.0000	188.000	357.0000	3970.0	
...	...	...	...	...	...	...	...	
prices[739]	194.000	1.600000	101.0000	31.0000	193.000	361.0000	3826.0	
prices[740]	208.000	1.600000	103.0000	41.0000	207.000	381.0000	4241.0	
prices[741]	197.000	1.600000	102.0000	29.0000	197.000	362.0000	4091.0	
prices[742]	185.000	1.600000	101.0000	16.0000	185.000	349.0000	4038.0	
prices[743]	174.000	1.700000	102.0000	7.0000	174.000	346.0000	3798.0	

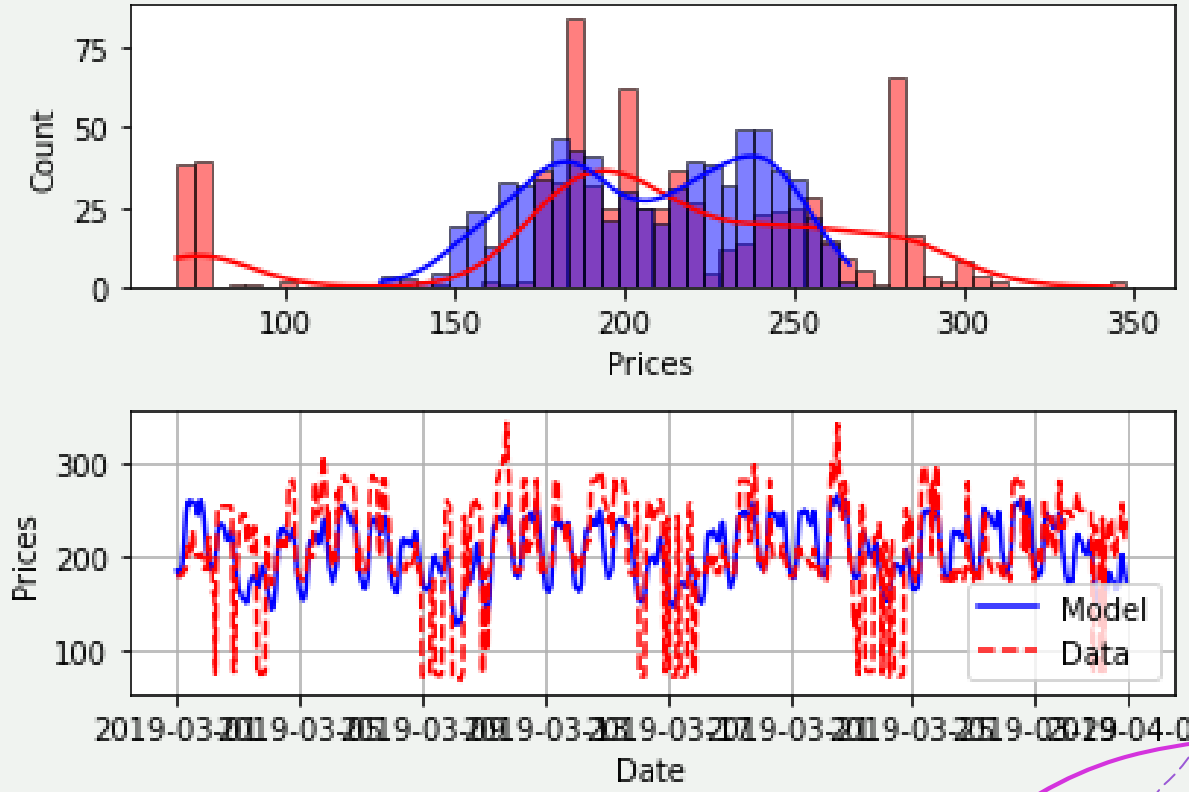
name	N_Eff/s	R_hat
lp__	NaN	NaN
a_demand	810.0	1.0
a_generation	750.0	1.0
offset_	780.0	1.0
prices[1]	786.0	1.0
...	...	...
prices[739]	757.0	1.0
prices[740]	839.0	1.0
prices[741]	810.0	1.0
prices[742]	799.0	1.0
prices[743]	752.0	1.0



Model 1 – Prices - Posterior

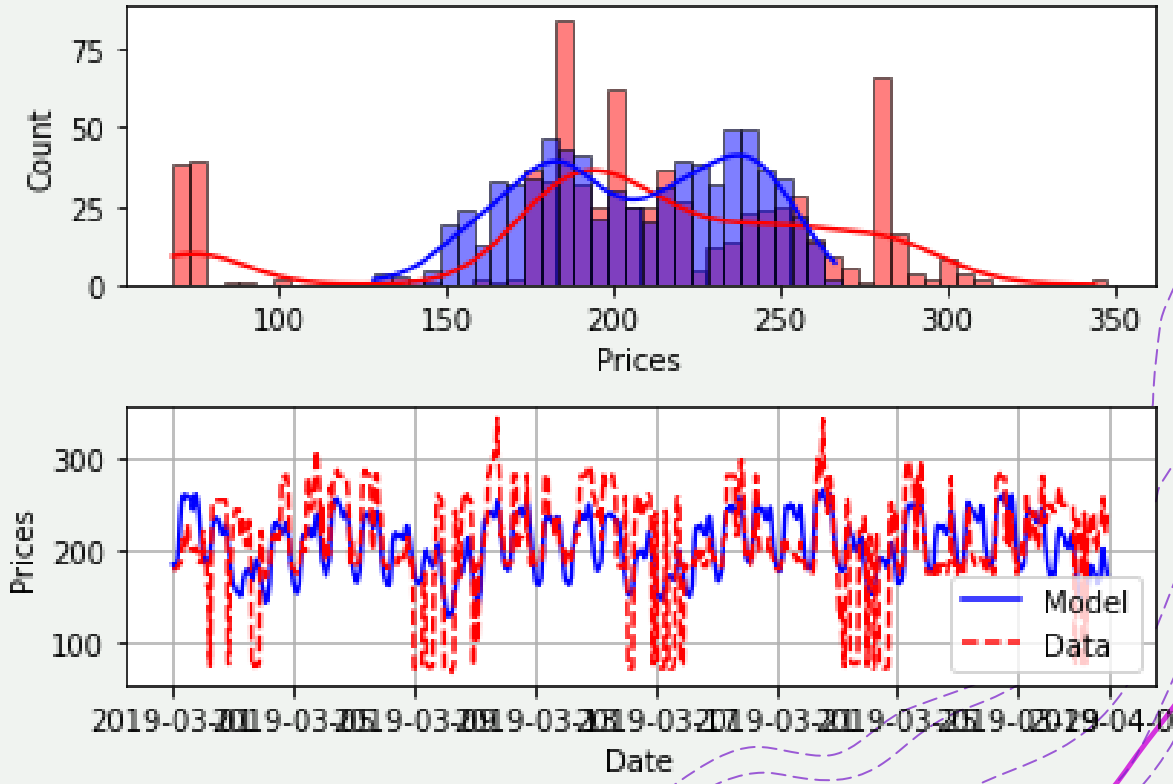
+Posterior model gave us results shown below

+The error was: RMSE = 48.14 / (max possible) 214.47



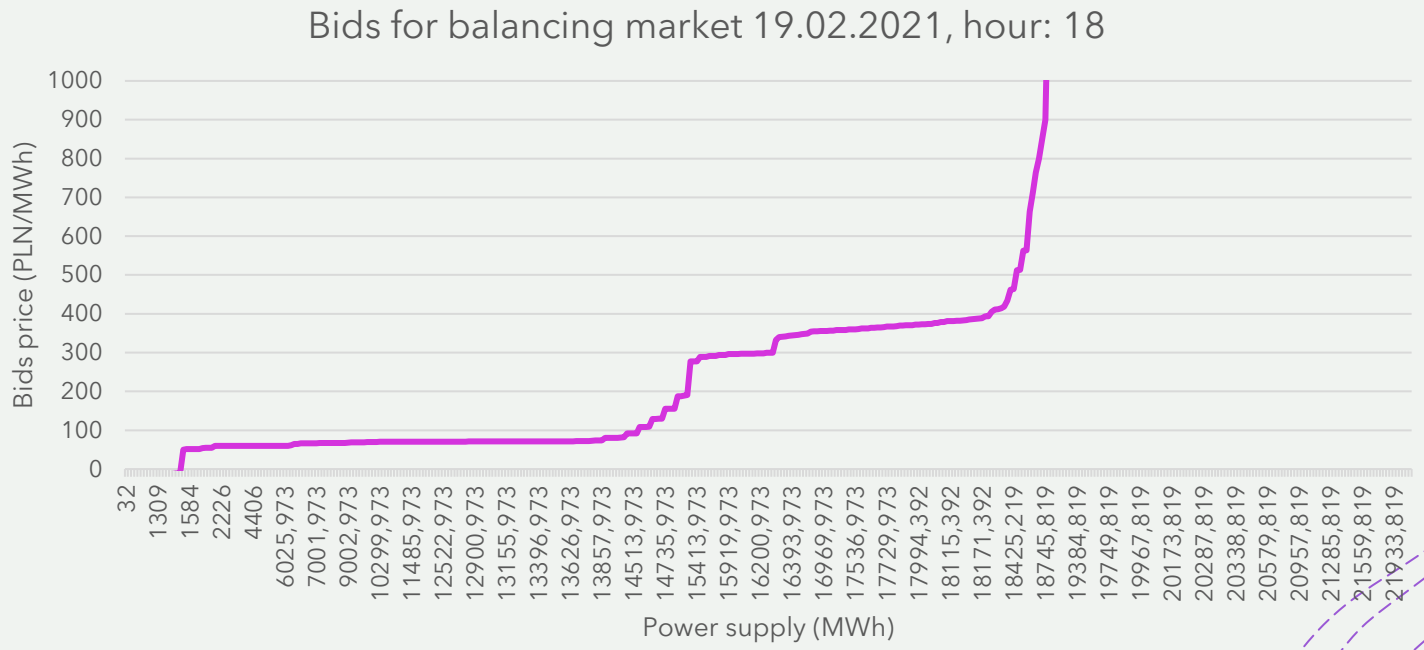
Model 1 – Prices – Posterior analysis

- Posterior predictive samples are quite consistient with data in middle range values, however they are not consistient with peaks in low and high values range



Model 1 – Prices – Posterior analysis

- Energy price data is very difficult to model because of price jumps (as shown on example data on the graph below)
- The jumps occur on different level of Residual load in different days

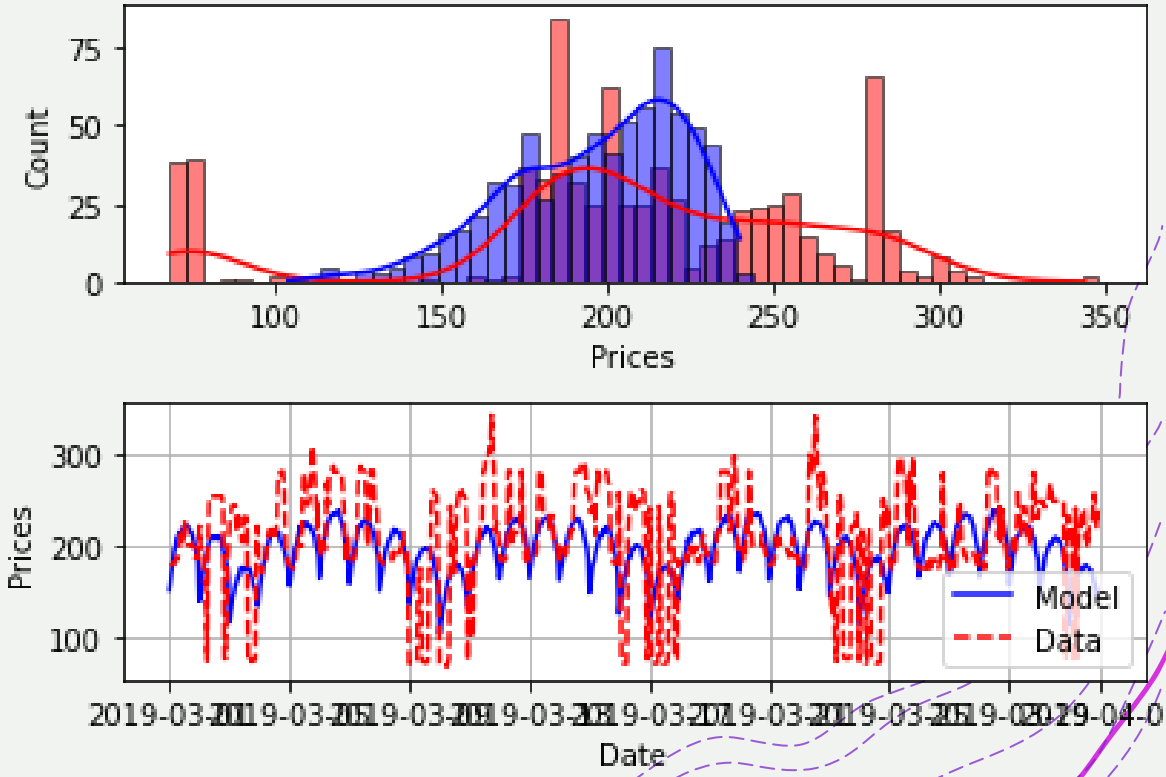


Model 1 – Overall posterior

- Overall posterior results are shown below
- The error value was: RMSE = 53.27 / (max possible) 214.47

Model 1 – Overall posterior analysis

- Overall postarior samples are somehow consistent with real data in middle values range
- However they do not cover spikes on extreme values
- Posterior samples values reach maximum of about 240, while real data reach up to 350

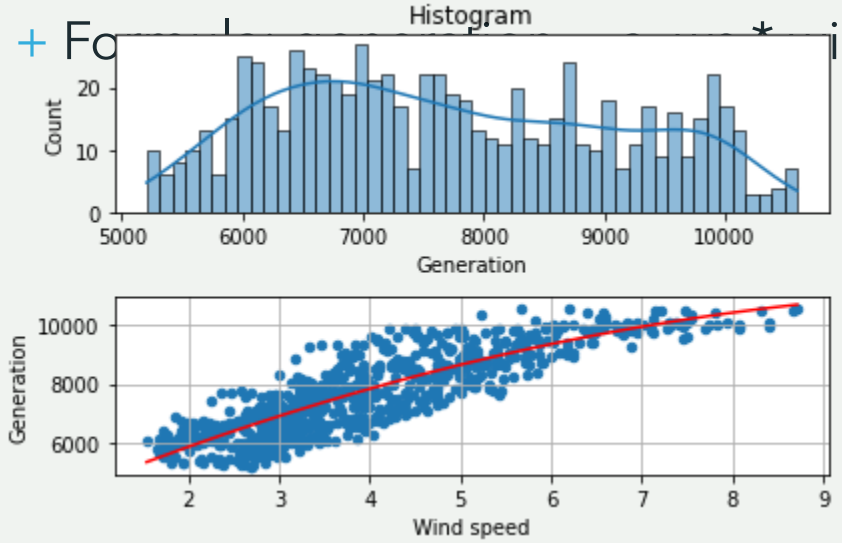


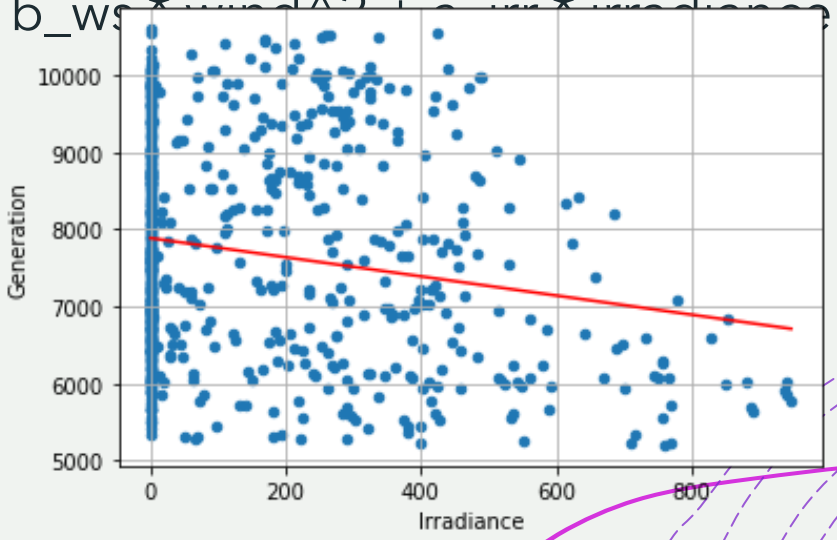
Model 2

- Demand prior and posterior for both models are the same
- 2nd model differs from the first one in generation – in the 2nd one the irradiance influence is calculated, while in the first it is omitted

Model 2 - Generation

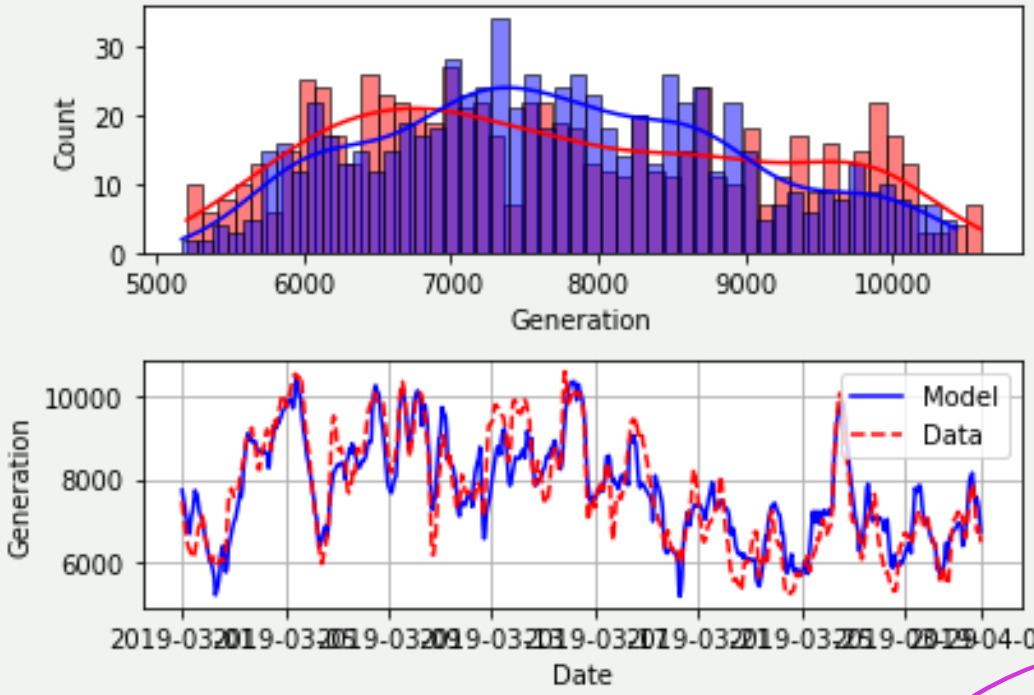
- Generation dependence on wind is the same as in the first model
- Wind speed has been fitted with quadratic function and irradiance with linear
- The irreadiance influence on generation is negative, which is a bit couterintuitive
- The reason for that can be explained with the fact that not only renewable Energy sources are included in non-controllable energy generation data, but also small conventional power plants
- Fo[illegible]ind + b_ws * wind^2 + a_irr * irradiance + offset_





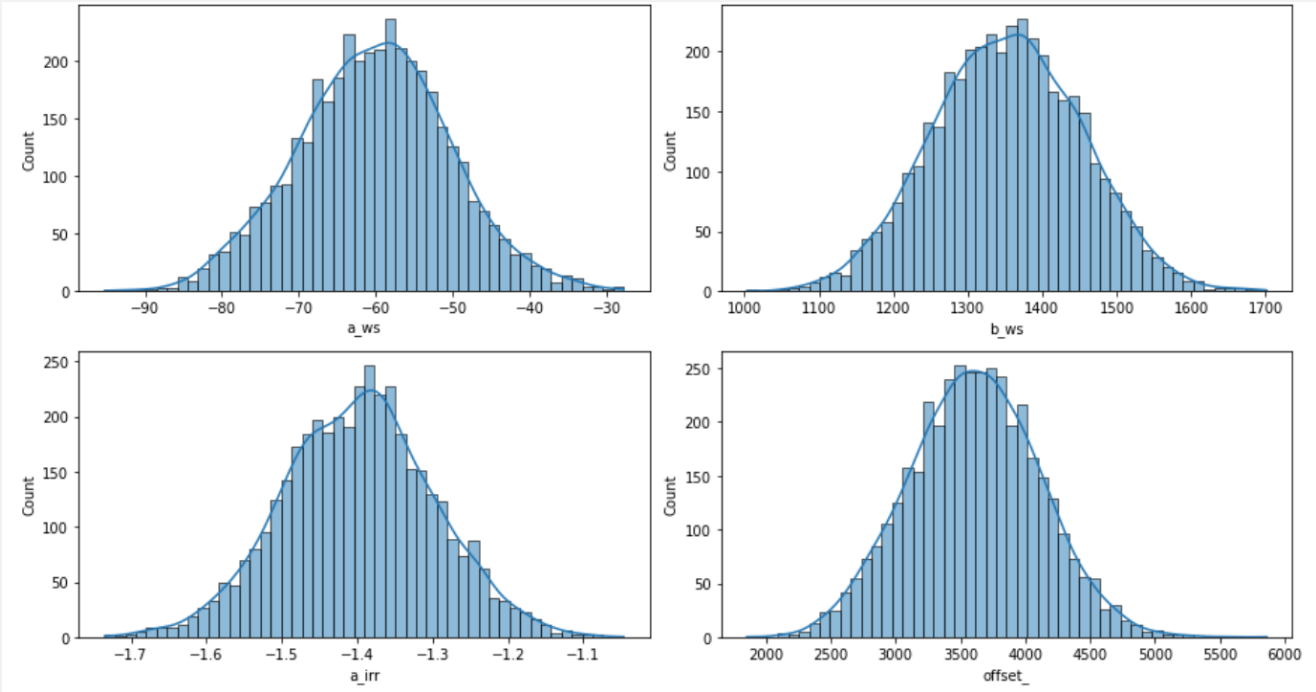
Model 2 – Generation - Prior

- Modelling prior in 2nd model (with irradiance) gave results shown on graph below
- The error value was: RMSE = 579.21 / (max possible) 7836.45



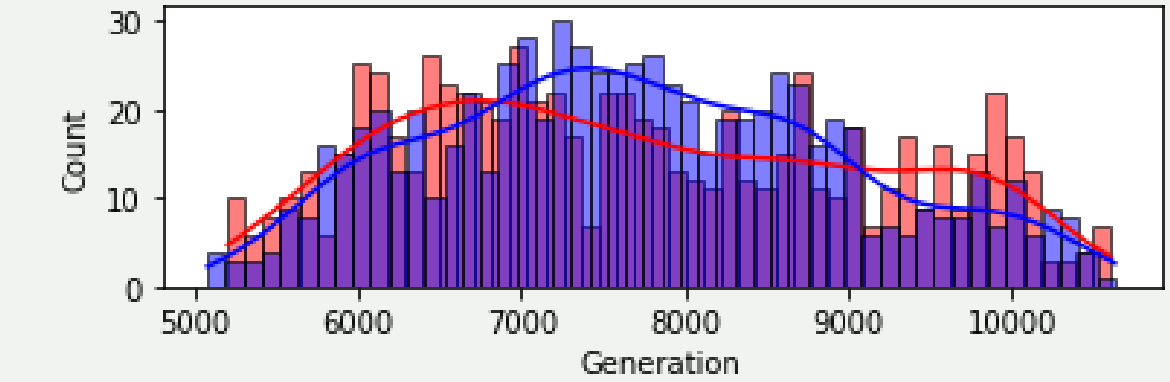
Model 2 – Generation – Prior summary

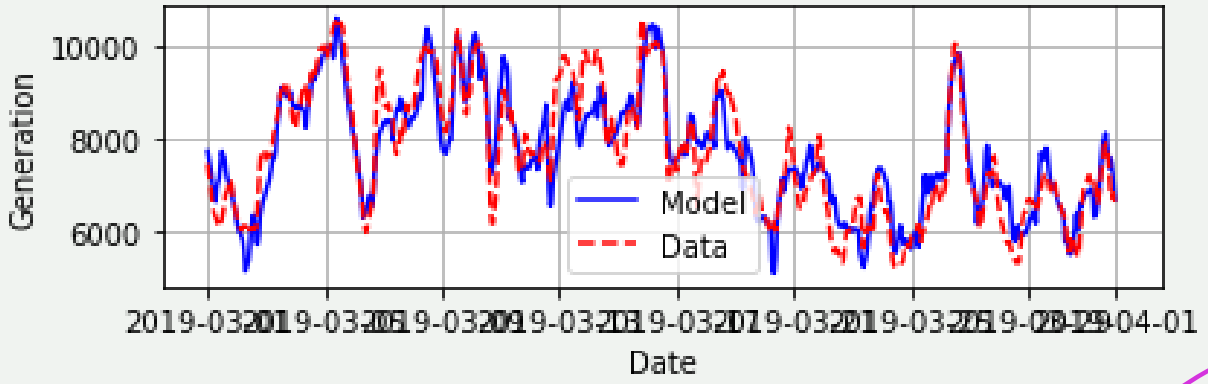
name	Mean	MCSE	StdDev	5%	50%	95%	N_Eff	N_Eff/s	R_hat
lp__	0.0	NaN	0.0	0.0	0.0	0.0	NaN	NaN	NaN
a_ws	-60.0	0.1500	9.9	-77.0	-60.0	-44.0	4100.0	580.0	1.0
b_ws	1400.0	1.6000	100.0	1200.0	1400.0	1500.0	4000.0	560.0	1.0
a_irr	-1.4	0.0016	0.1	-1.6	-1.4	-1.2	4000.0	570.0	1.0
offset_	3600.0	8.3000	500.0	2800.0	3600.0	4400.0	3700.0	520.0	1.0
...	...	...	...	...	...	...	...	...	...
generation[739]	7565.0	19.0000	1181.0	5614.0	7562.0	9530.0	3984.0	566.0	1.0
generation[740]	7455.0	19.0000	1193.0	5485.0	7458.0	9437.0	3828.0	544.0	1.0
generation[741]	7315.0	18.0000	1157.0	5430.0	7311.0	9198.0	3931.0	558.0	1.0
generation[742]	7033.0	18.0000	1147.0	5112.0	7026.0	8933.0	4074.0	579.0	1.0
generation[743]	6736.0	19.0000	1158.0	4817.0	6735.0	8651.0	3832.0	544.0	1.0



Model 2 – Generation - Posterior

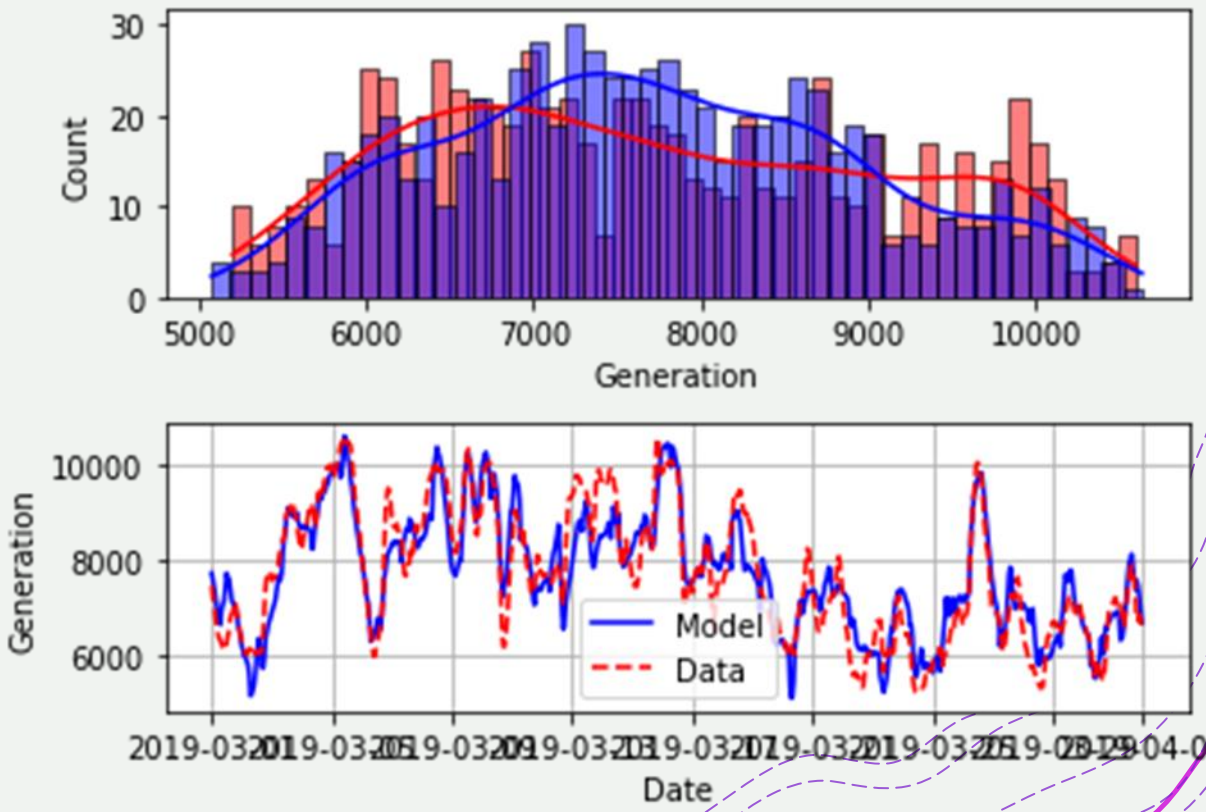
- Simulating posteriori gave results shown below
- The error value is: RMSE = 569.56 / (max possible) 7836.45





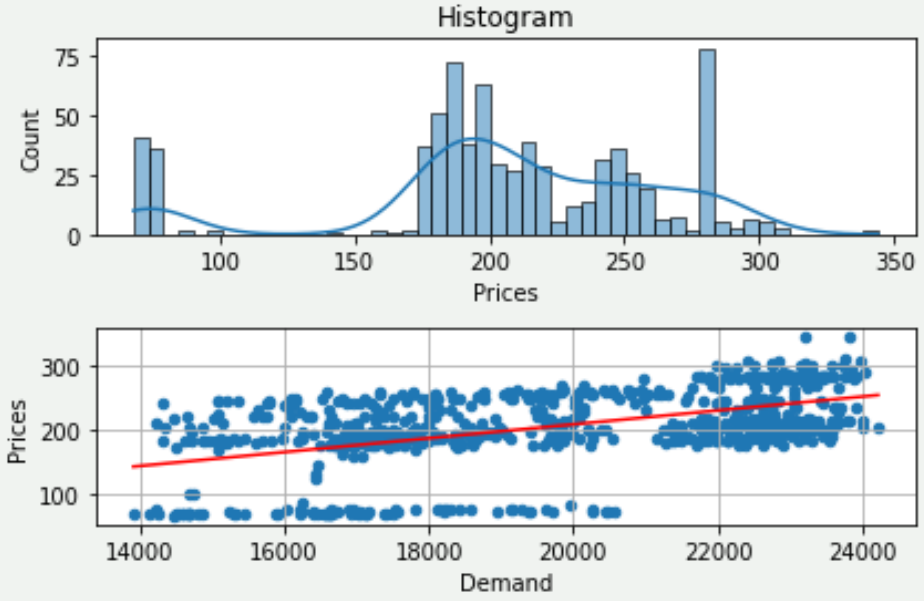
Model 2 – Generation – Posterior analysis

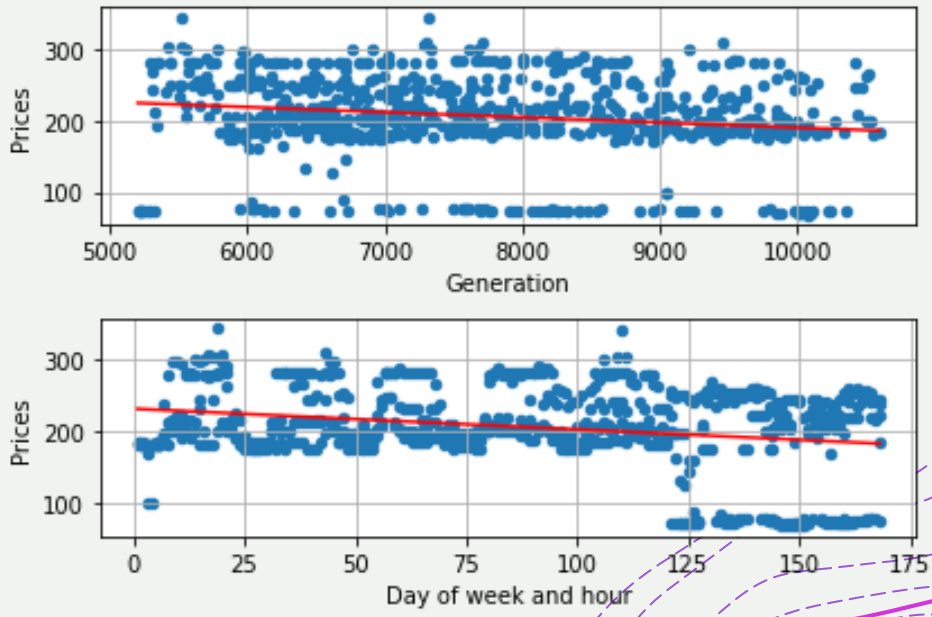
- Posterior predictive samples are consistent with real data
- There are a bit more samples in middle range of generation values



Model 2 - Prices

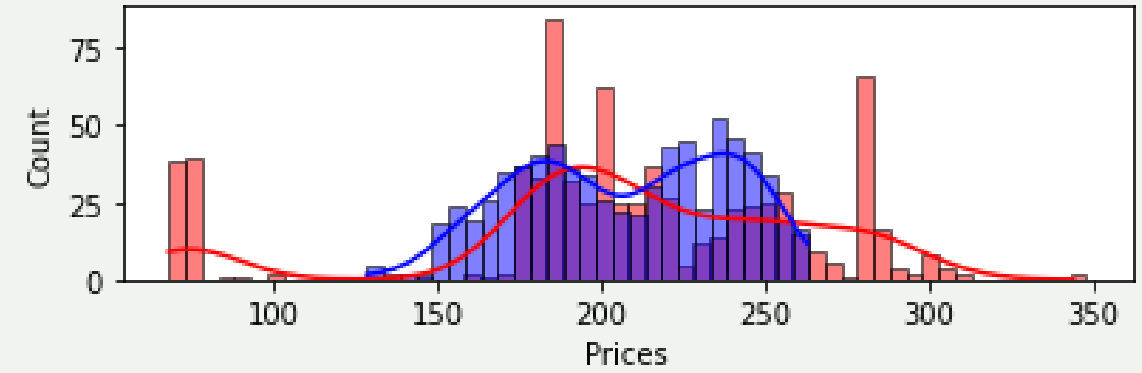
- Same as for the 1st model we have tried to determine the relations with load, generation and day of week with linear functions
  - Formula: prices = a_demand * demand + a_generation * generation + offset_

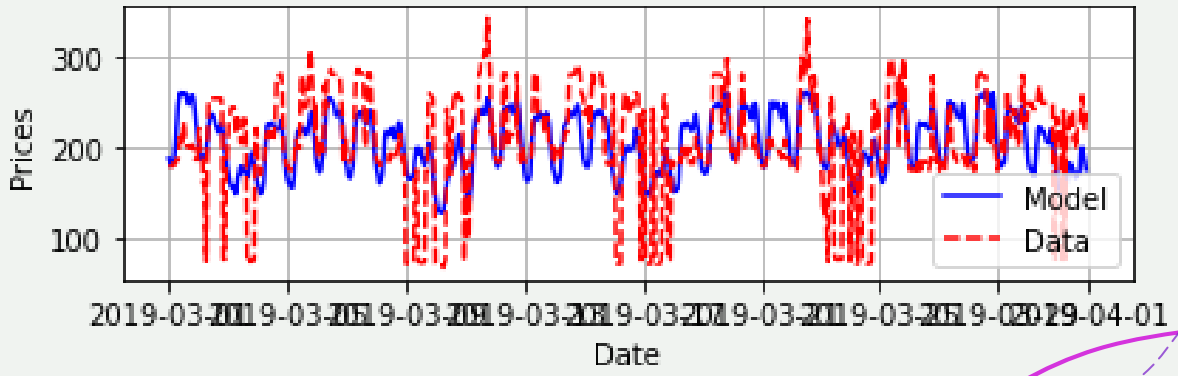




Model 2 – Prices - Prior

- Prior for the prices gave results shown below
- The error value was: RMSE = 48.13 / (max possible) 214.47

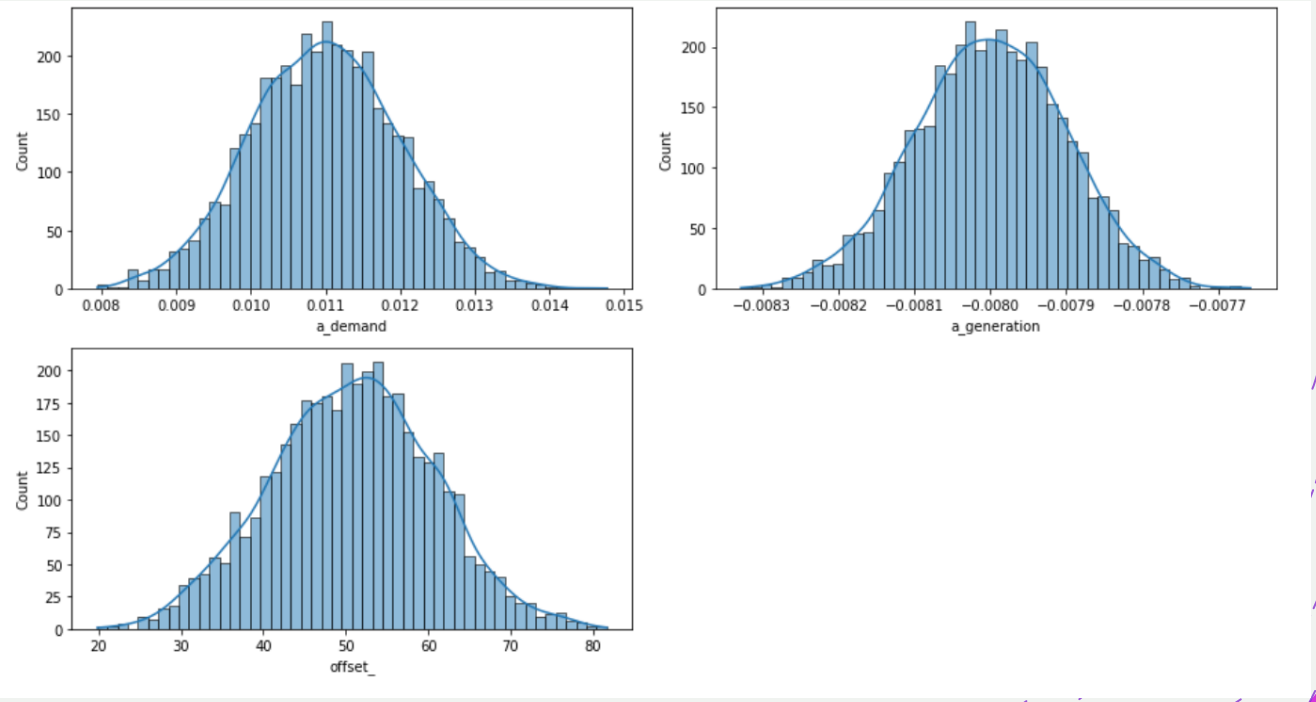




Model 2 – Prices – Prior summary

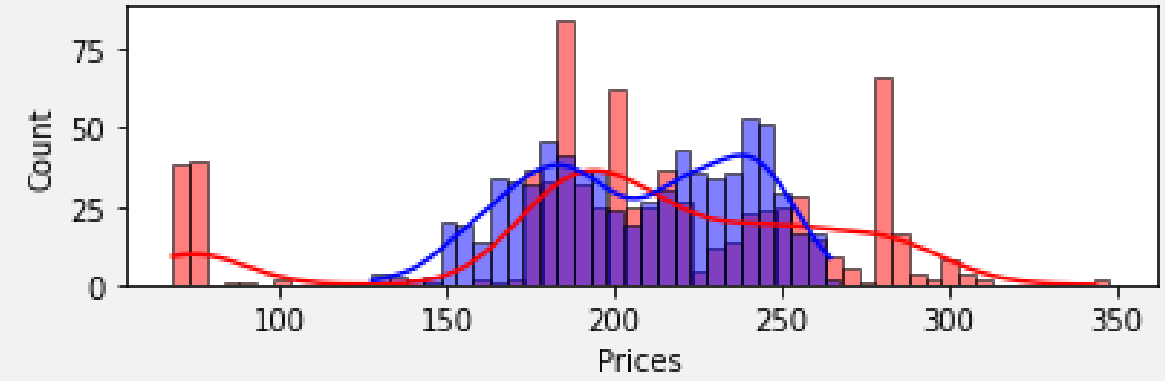
name	Mean	MCSE	StdDev	5%	50%	95%	\
lp__	0.000	NaN	0.000000	0.0000	0.000	0.0000	
a_demand	0.011	0.000016	0.001000	0.0094	0.011	0.0130	
a_generation	-0.008	0.000002	0.000099	-0.0082	-0.008	-0.0078	
offset_	51.000	0.160000	9.900000	34.0000	51.000	67.0000	
prices[1]	185.000	1.700000	102.000000	15.0000	184.000	351.0000	
...	...	...	...	...	...	...	
prices[739]	194.000	1.700000	103.000000	25.0000	192.000	365.0000	
prices[740]	204.000	1.700000	104.000000	31.0000	204.000	373.0000	
prices[741]	190.000	1.700000	102.000000	26.0000	190.000	362.0000	
prices[742]	181.000	1.700000	101.000000	17.0000	179.000	348.0000	
prices[743]	177.000	1.600000	101.000000	6.9000	177.000	342.0000	

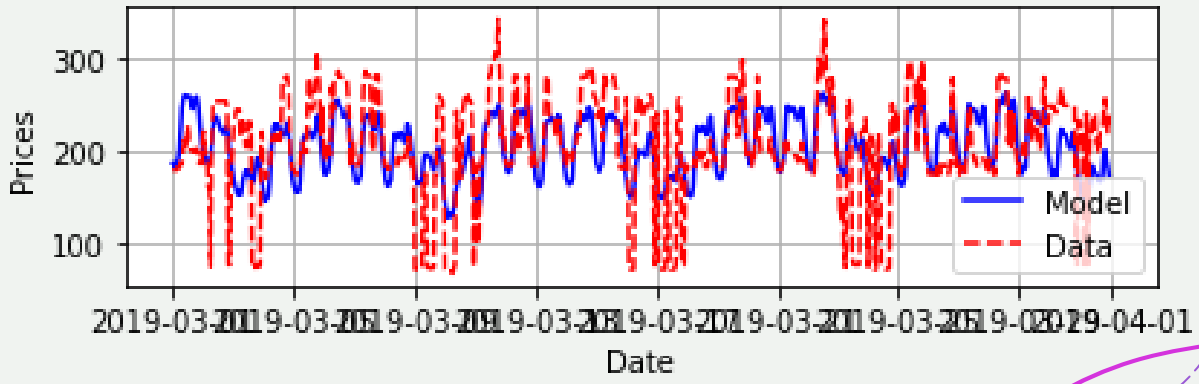
name	N_Eff	N_Eff/s	R_hat
lp__	NaN	NaN	NaN
a_demand	4100.0	870.0	1.0
a_generation	4000.0	840.0	1.0
offset_	4000.0	840.0	1.0
prices[1]	3741.0	784.0	1.0
...	...	...	...
prices[739]	3836.0	803.0	1.0
prices[740]	3753.0	786.0	1.0
prices[741]	3805.0	797.0	1.0
prices[742]	3725.0	780.0	1.0
prices[743]	4066.0	852.0	1.0



Model 2 – Prices – posterior

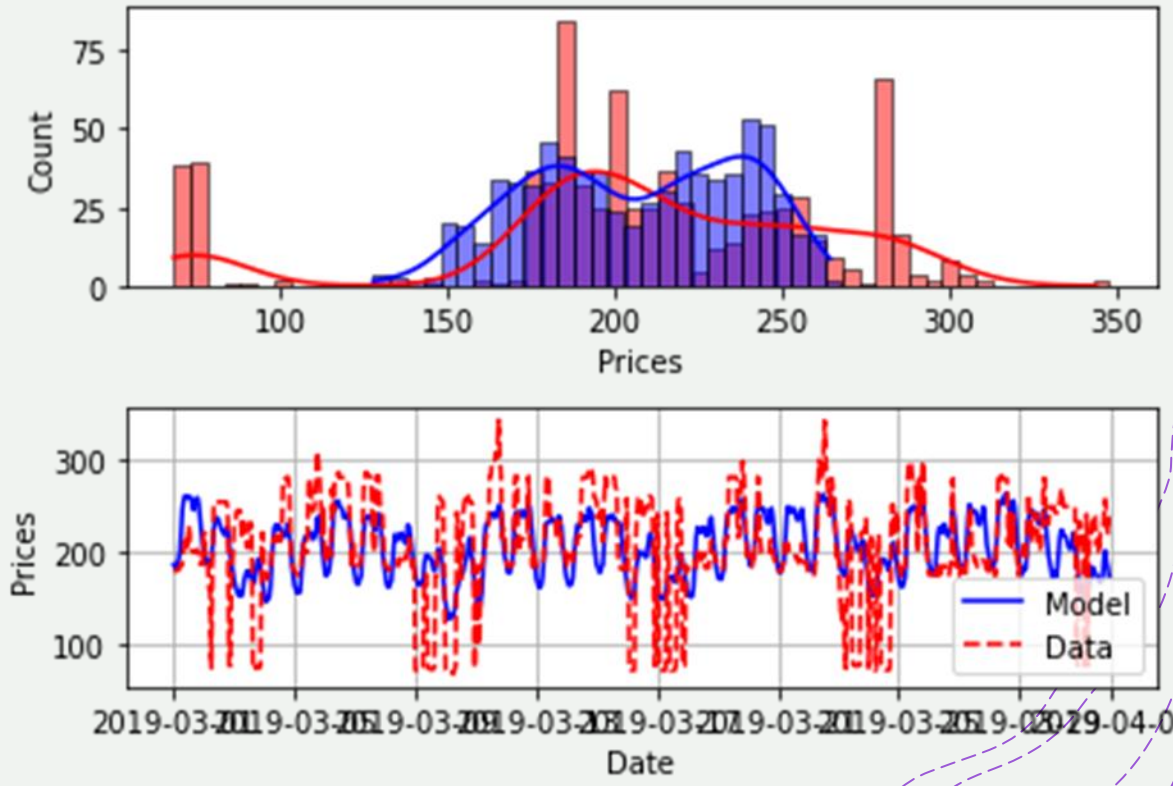
- Posterior simulation for prices gave results shown below
- The error value was: RMSE = 48.11 / (max possible) 214.47





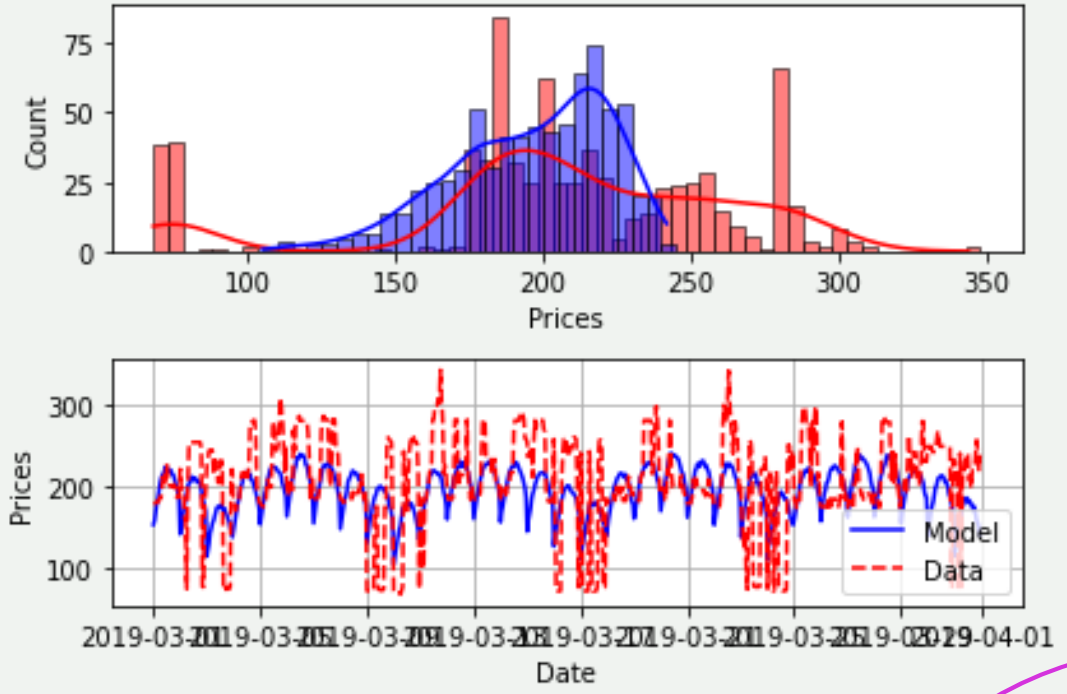
Model 2 – Prices – posterior analysis

- Same as for model 1 posterior predictive samples are quite consistient with data in middle range values, however they are not consistient with peaks in low and high values range



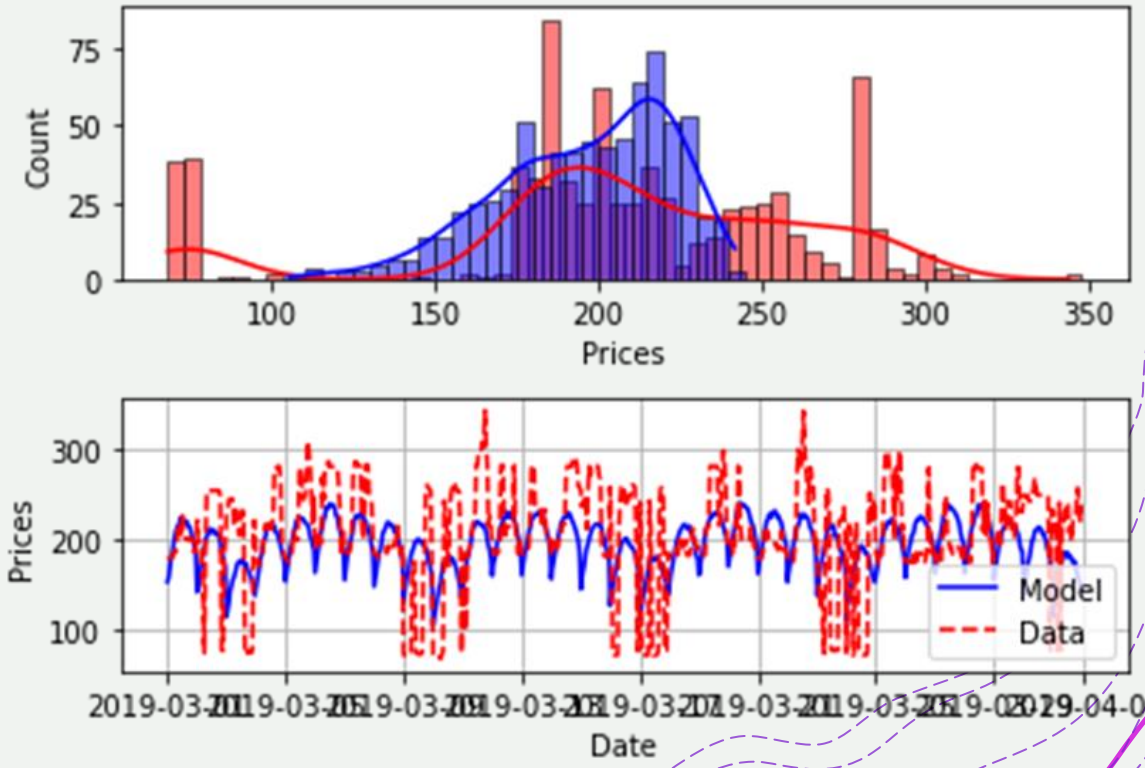
Model 2 – Overall posterior

- Overall posteriori simulation for model 2 gave results shown below
- The error value was: RMSE = 53.86 / (max possible) 214.47



Model 2 – Overall posterior analysis

- Model 2 behaves really similarly to model 1
- Overall postarior samples are somehow consistent with real data in middle values range
- However they do not cover spikes on extreme values
- Posterior samples values reach maximum of about 240, while real data reach up to 350



Model comparission

+Predictions for 2020

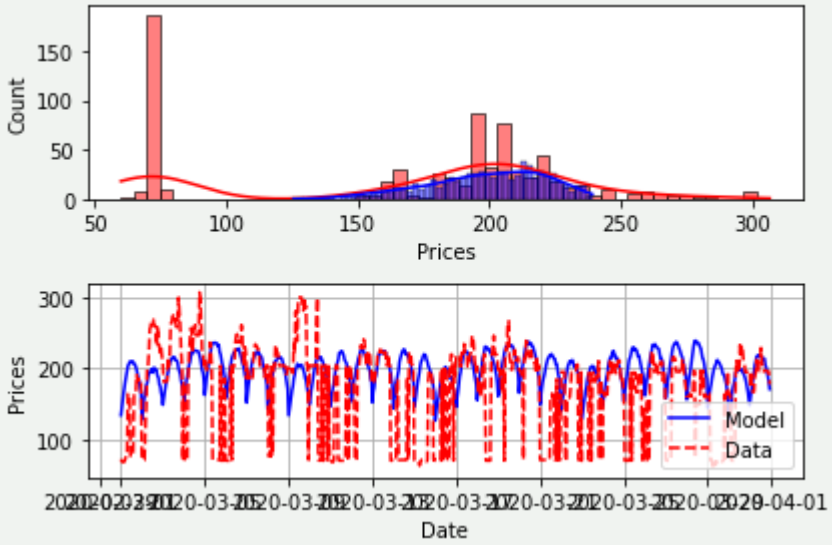
Model with irradiance

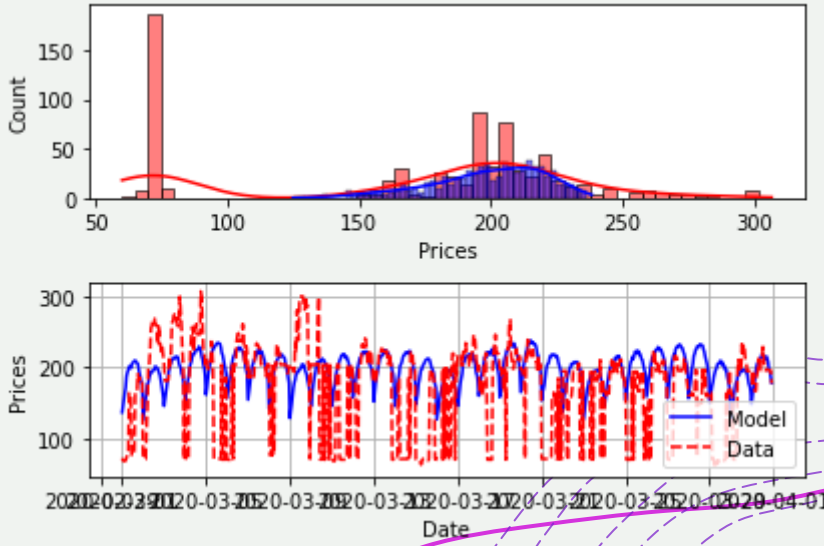
RMSE = 70.40 / (max possible) 180.57

Model without irradiance

RMSE = 69.84 / (max possible) 180.57

+ Both models give similar results, model without irradiance has a bit lower RMSE



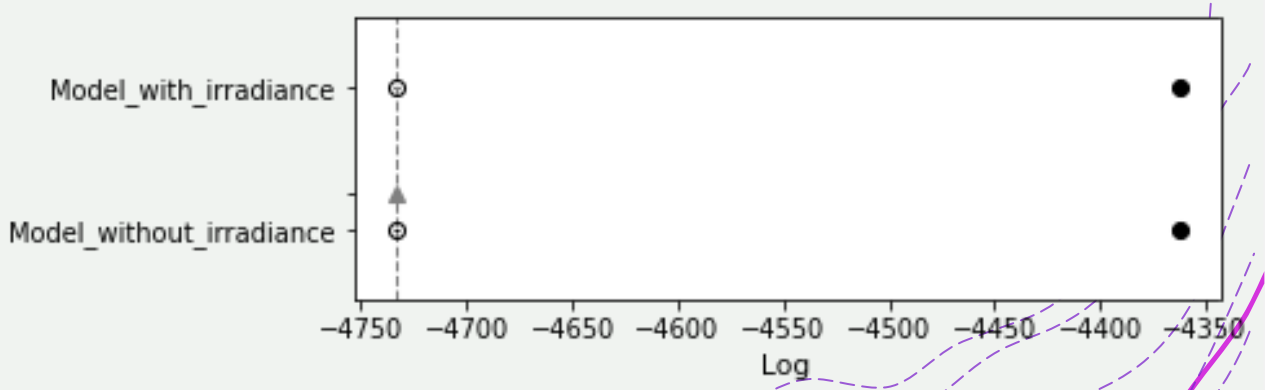


Model comparision using information criteria - waic

- As shown on waic comparisson graph and statistics both models perform really simillar, which is because the only difference between tchem is taking into calculations dependence on solar irradiation which doas not have big influence on final data

	rank	waic	p_waic	d_waic	weight	\
Model_with_irradiance	0	-4733.280382	371.314863	0.00000	0.654092	
Model_without_irradiance	1	-4733.518211	371.251251	0.23783	0.345908	

	se	dse	warning	waic_scale
Model_with_irradiance	0.872827	0.000000	True	log
Model_without_irradiance	0.926194	1.242111	True	log

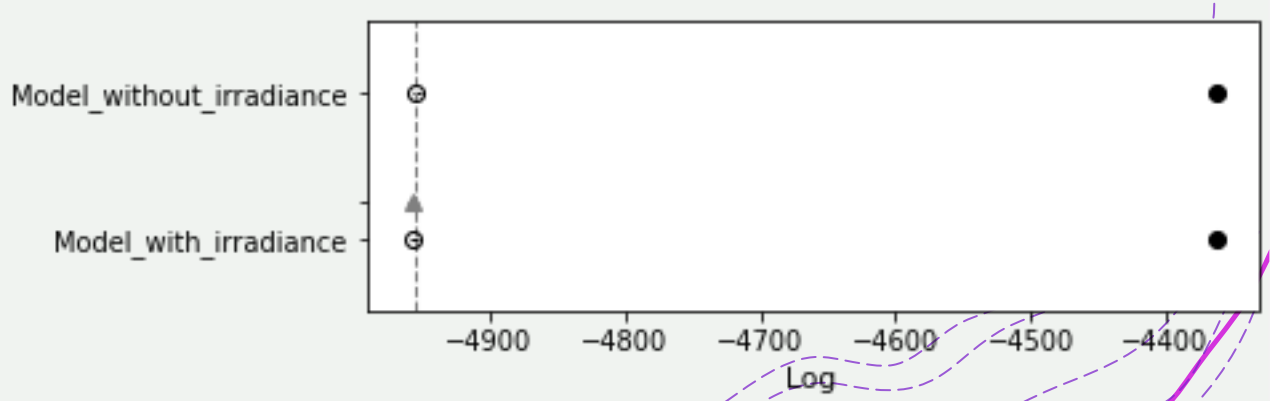


Model comparision using information criteria - loo

- Same as for waic, in comparision using loo the results are also very simmilar
- Besed on the information given by waic and loo criteria Model 1 and Model 2 overlap

	rank	loo	p_loo	d_loo	weight	\
Model_without_irradiance	0	-4954.314111	592.047150	0.000000	0.618524	
Model_with_irradiance	1	-4956.698076	594.732556	2.383965	0.381476	

	se	dse	warning	loo_scale
Model_without_irradiance	3.243580	0.000000	True	log
Model_with_irradiance	3.198381	4.489707	True	log

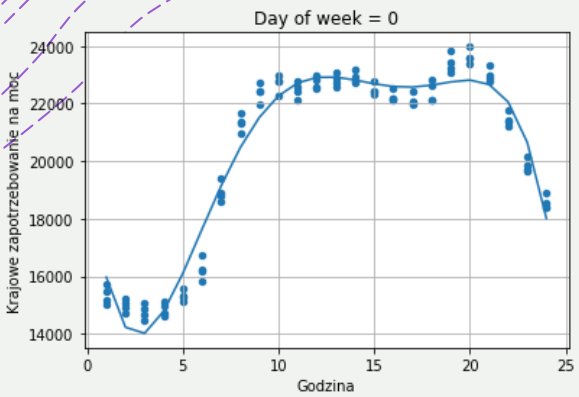


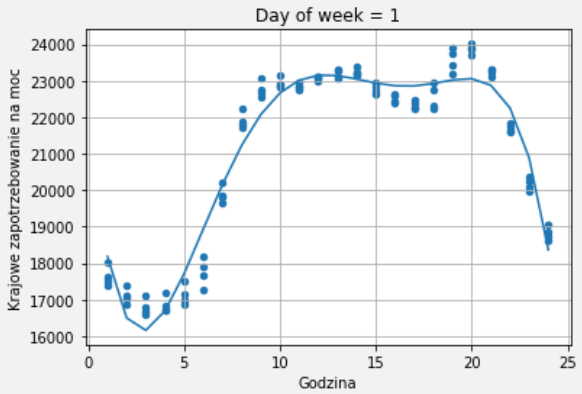
Additional model

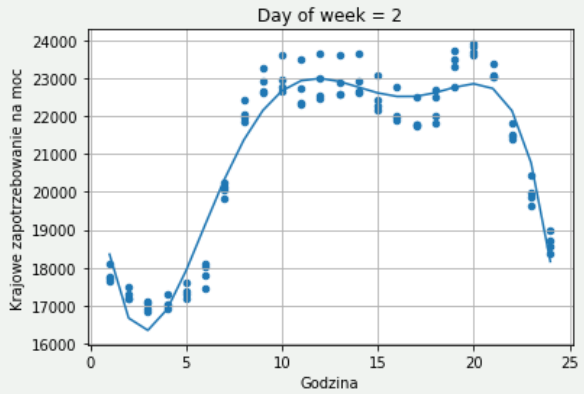
- Additional model was based on fitting distributions of data with polynomial regressions
- No histograms were used
- Priors were based on student and normal distributions

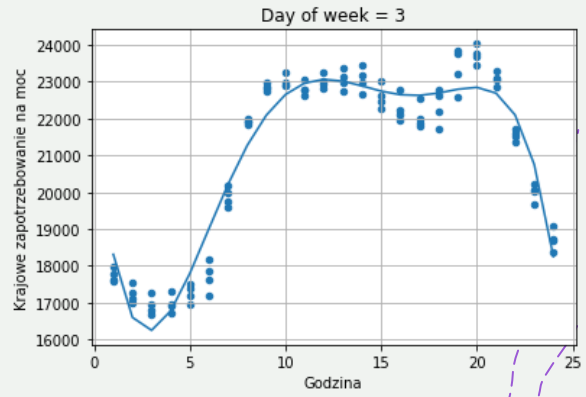
Additional model - Demand

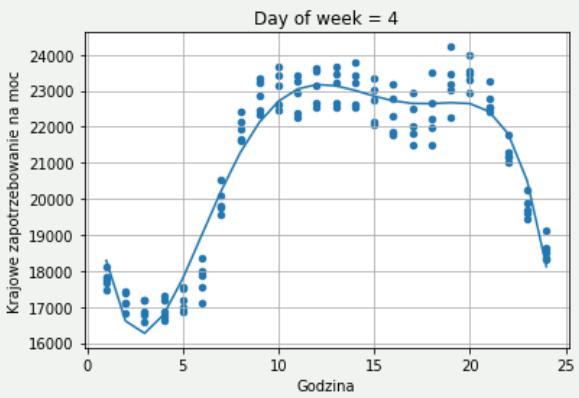
- Hourly demand distriution has been done separately for each weekday
- Distributions are fitted with 5th degree polynomial

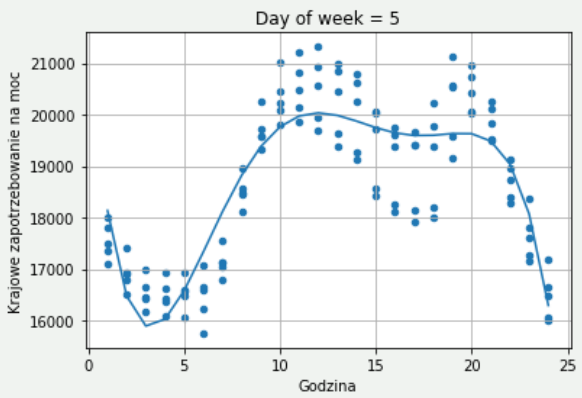


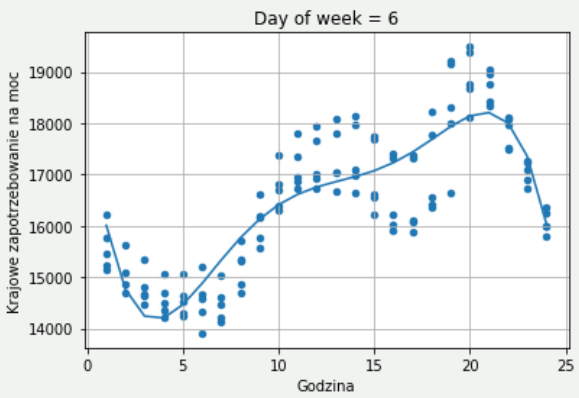


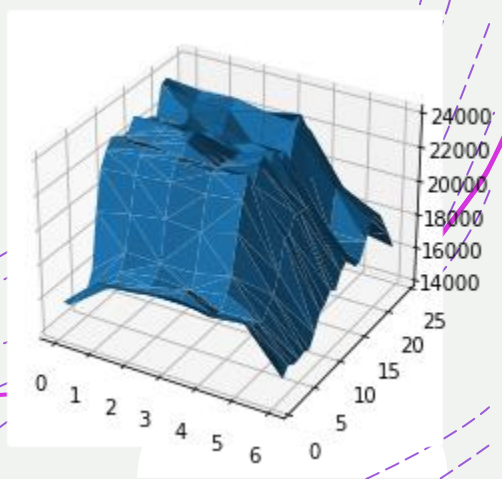






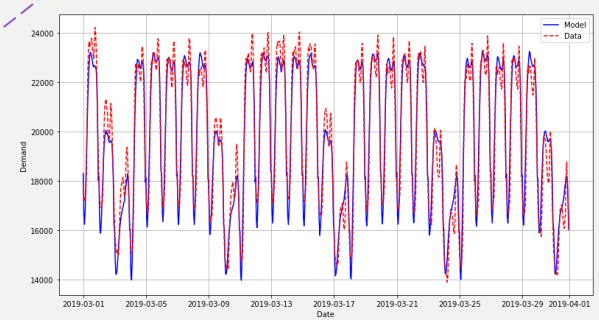


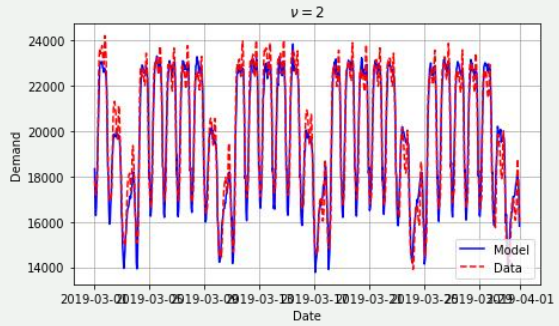


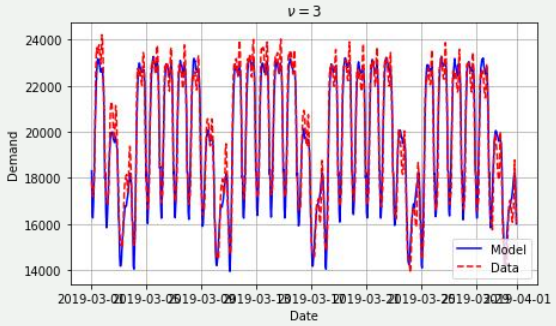


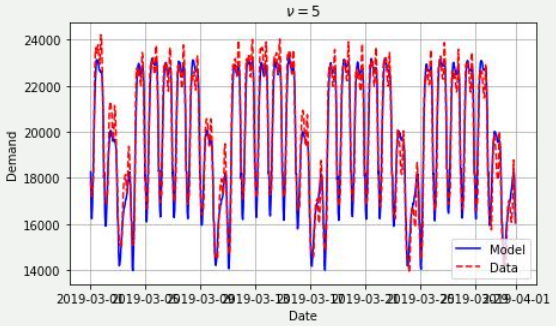
Additional model – Demand - prior

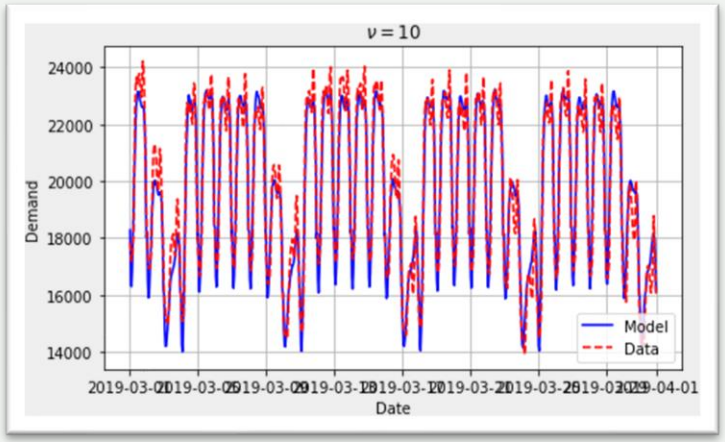
- We have tested normal distribition and T-student distributions with various degrees of freedom
- The best results were reached for nu=10.

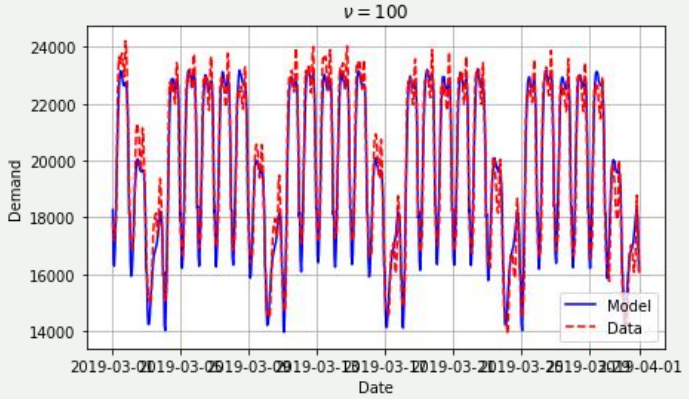


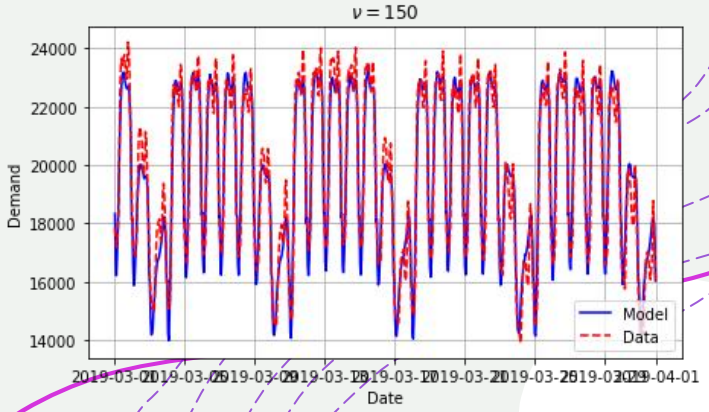








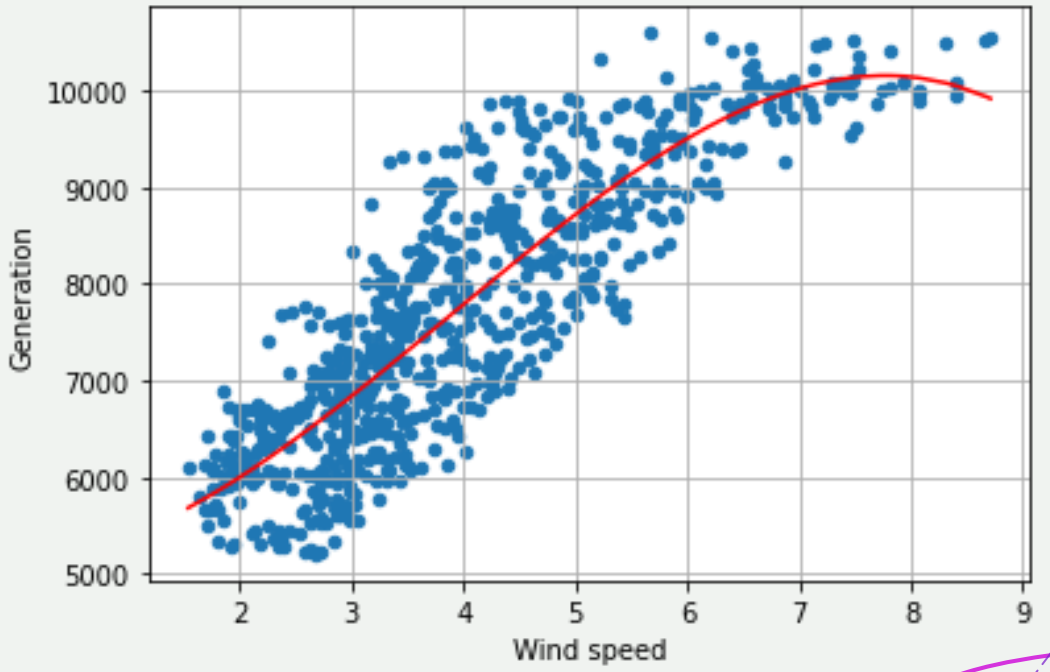




Additional model –Generation

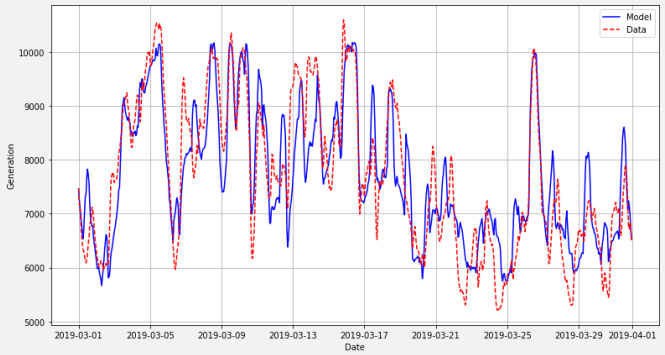
+As for the 1st model the influence of irradiation has been omitted

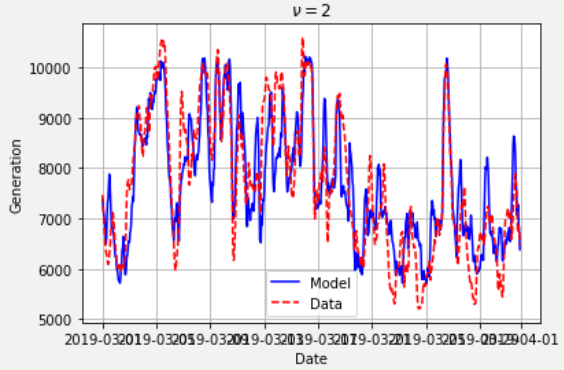
+The wind influence was fitted with 3rd degree polynomial

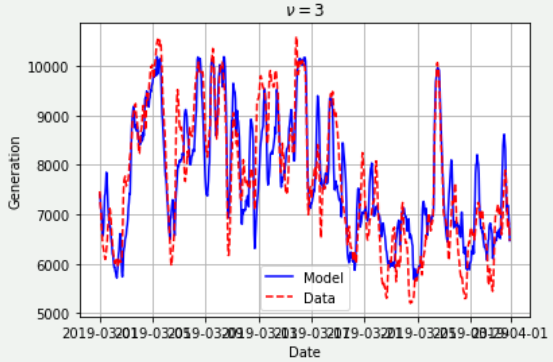


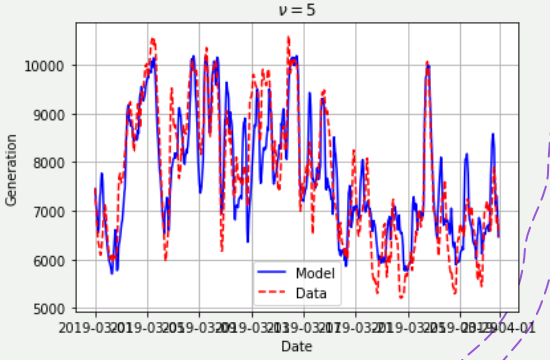
Additional model – Generation – prior

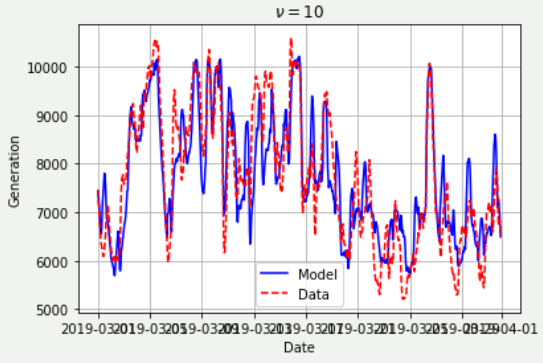
- We have tested normal distribition and T-student distributions with various degrees of freedom
- The best results were reached for nu=100.

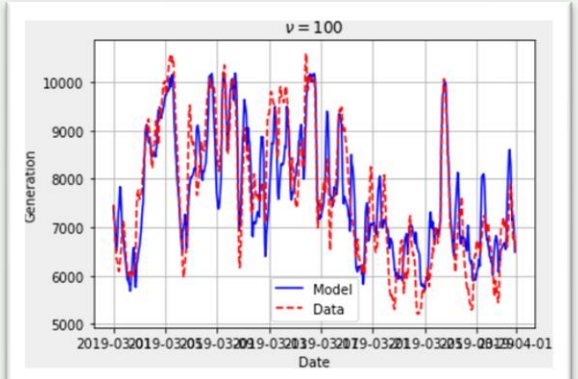


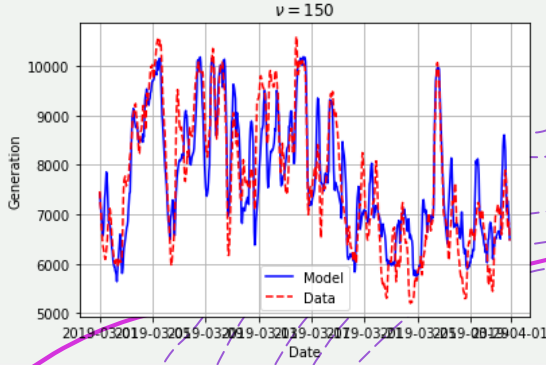






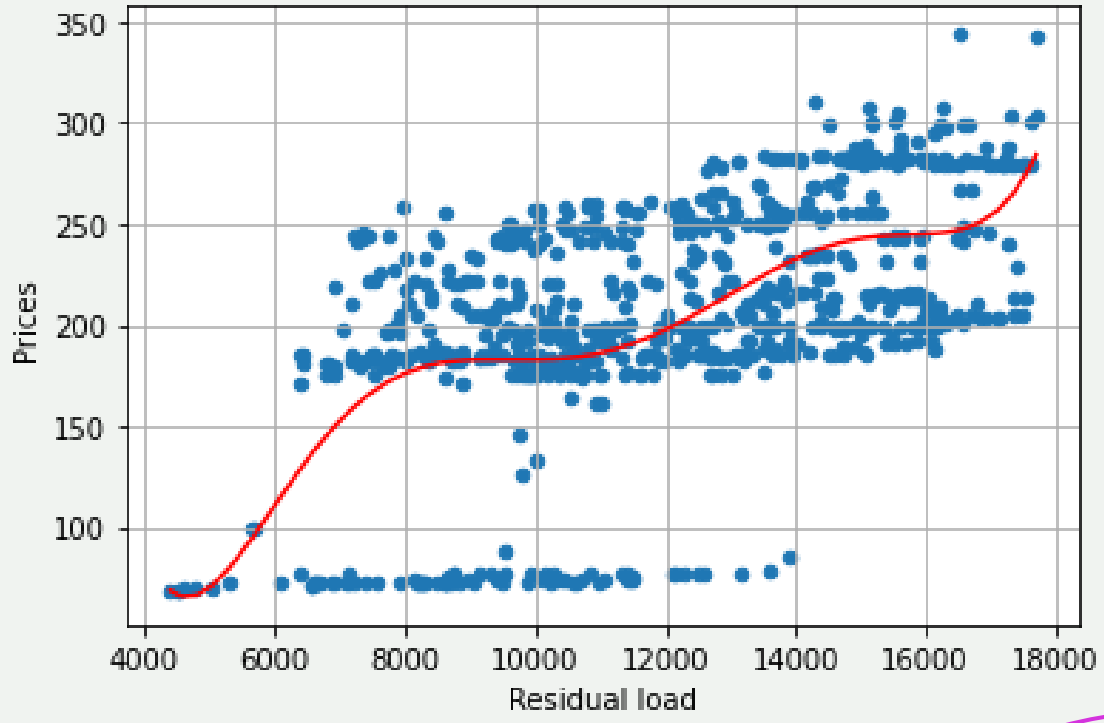






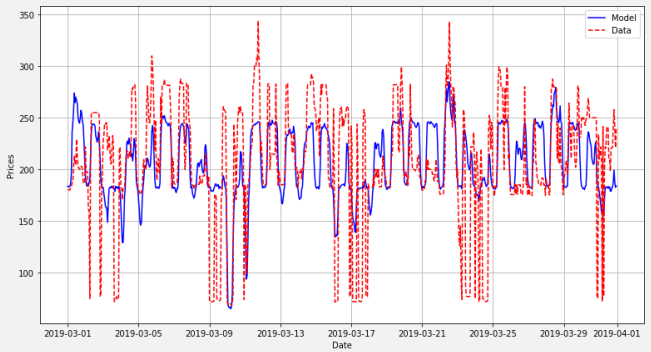
Additional model – prices

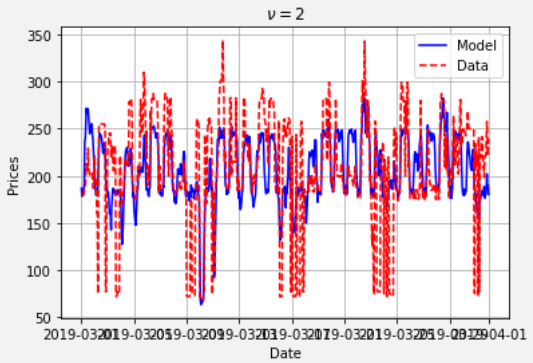
+ The dependence of prices on the residual load has been fitted with 5th degree polynomial

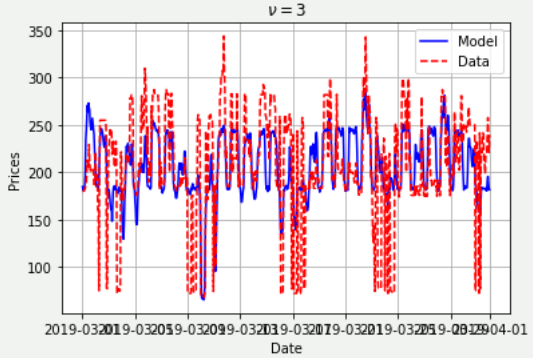


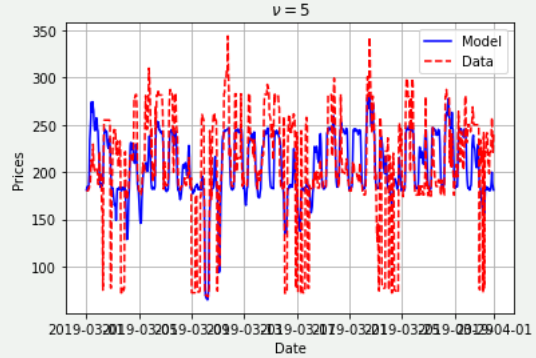
Additional model – prices - prior

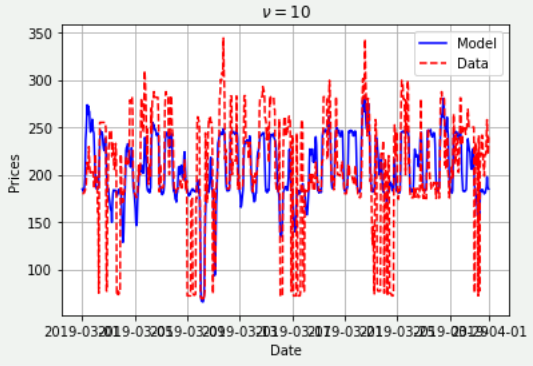
- We have tested normal distribition and T-student distributions with various degrees of freedom
- The best results were reached for nu=100.

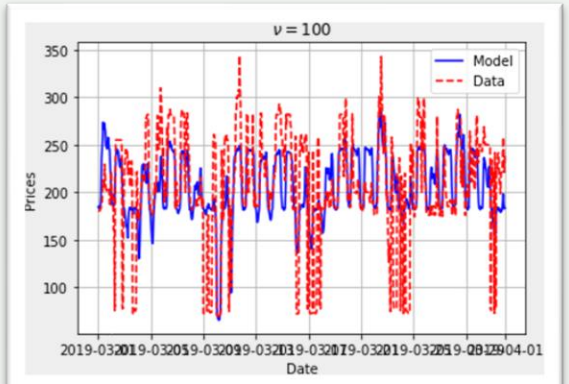


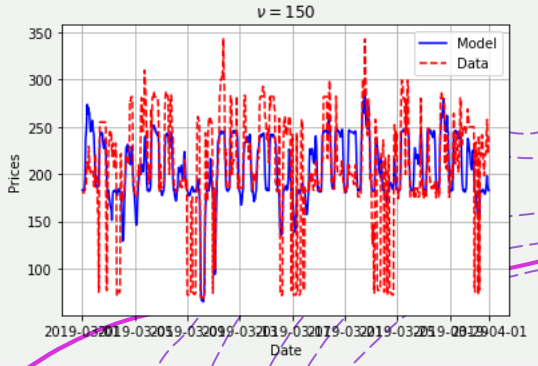












Additional model – Posterior

- Posterior for the additional model gaze the results shown on the graph below
- The error value was: RMSE = 50.57 / (max possible) 214.47
- The posterior samples are quite consistent with real data but they do not reach the extremes of real prices