Report on learning practice # 1

Analysis of univariate random variables

Performed by:

Patrik Stronski

Ivan Pavelev

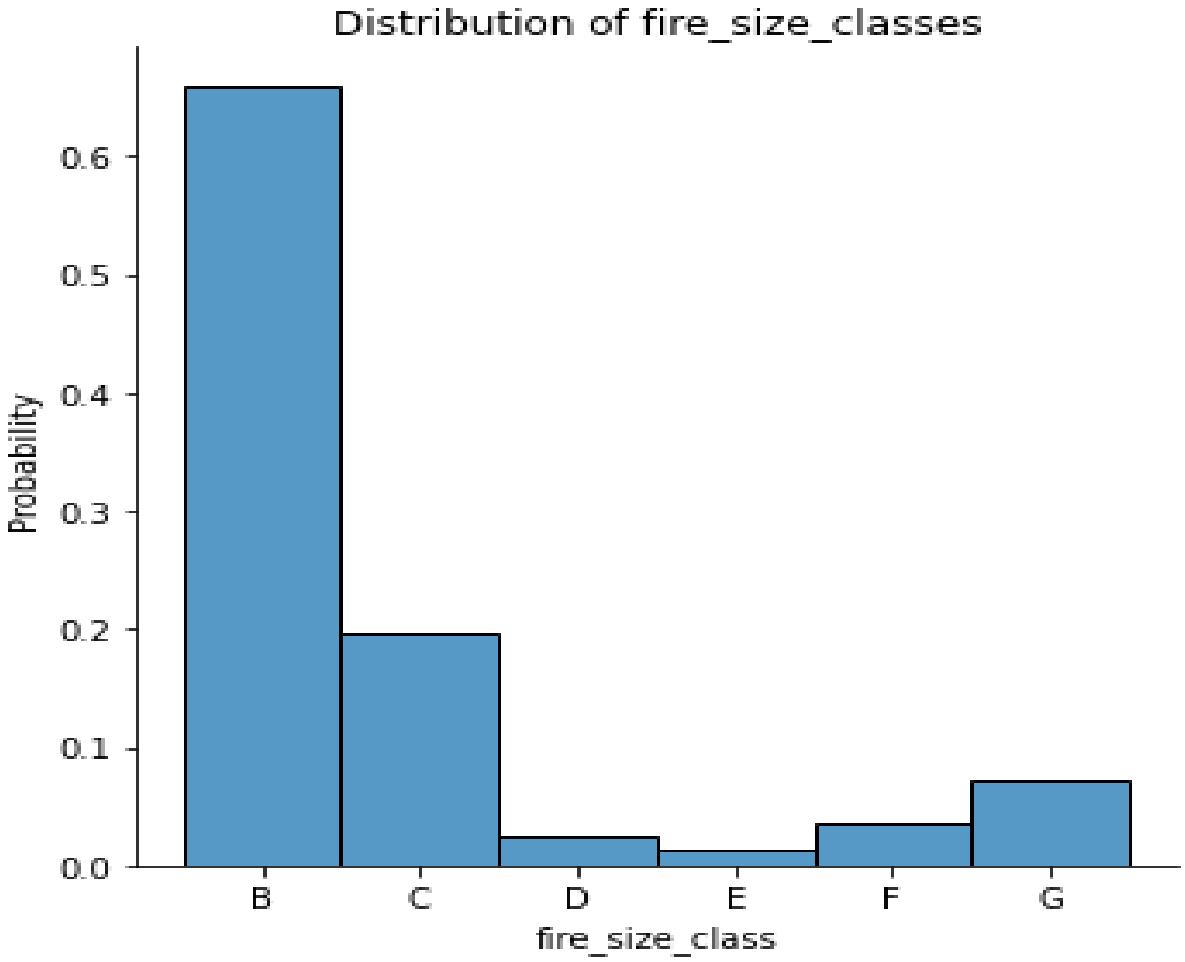
J4133C

Saint-Petersburg

2021

0. Dataset description:

The dataset we used collects the data about fires in the USA. It is a subset of the bigger dataset. The dataset is contained within one CSV file, easy to read and process. The main variable in the dataset is *fire_size* which presents how big (in acres) the fire was. The fires are divided into several categories based on their size – *fire_size_class*. The classes possible here are from A to G, however in the dataset we used no small fires (A-class) are contained.



The dataset we used: https://www.kaggle.com/capcloudcoder/us-wildfire-data-plus-other-attributes?select=FW_Veg_Rem_Combined.csv

The base dataset: https://www.kaggle.com/rtatman/188-million-us-wildfires

1. Substantiation of chosen subsample;

We have chosen to assess 4 variables from our dataset:

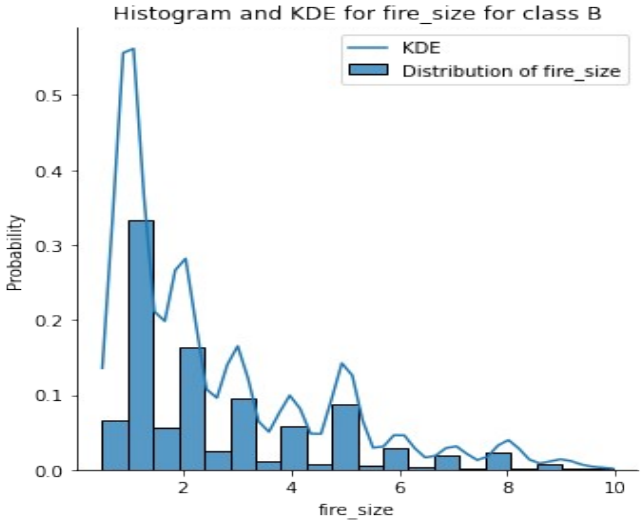
- fire_size — size of fire (in acres)
- Temp_pre_7 — average temperature 7 days before the fire was discovered (in degress Celsius)
- Hum_pre_7 — average humidity 7 days before the fire was discovered (in %)

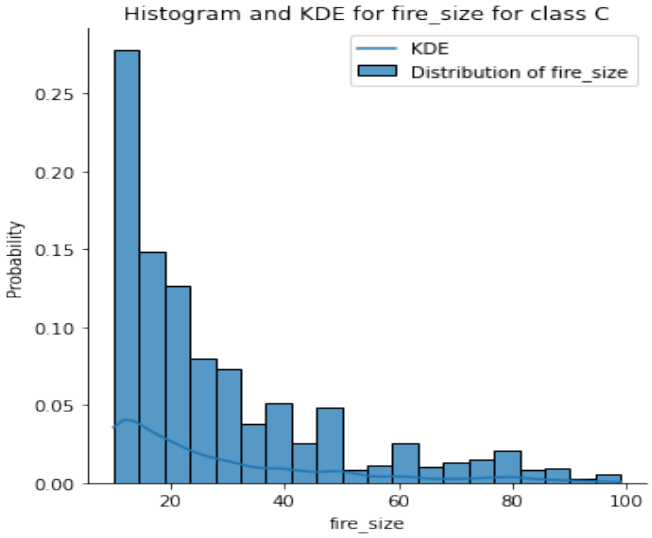
- Wind_pre_7 — average wind speed 7 days before the fire was discovered (in m/s)

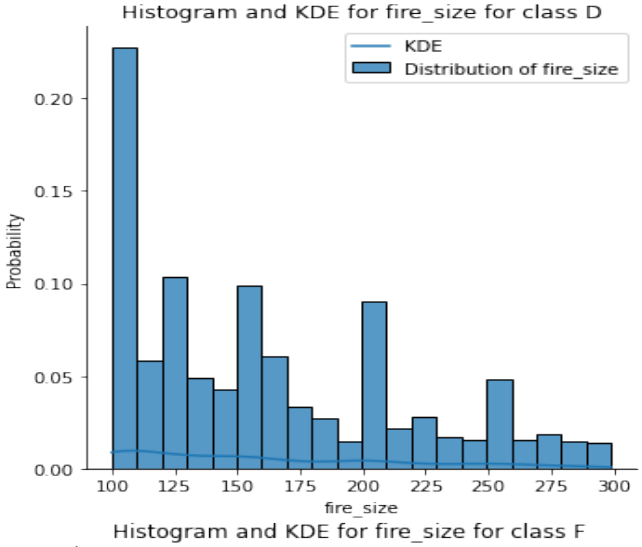
Fire_size we model for each class (how big is the fire per some class), whereas temperature, humidity, wind we model using the whole dataset.

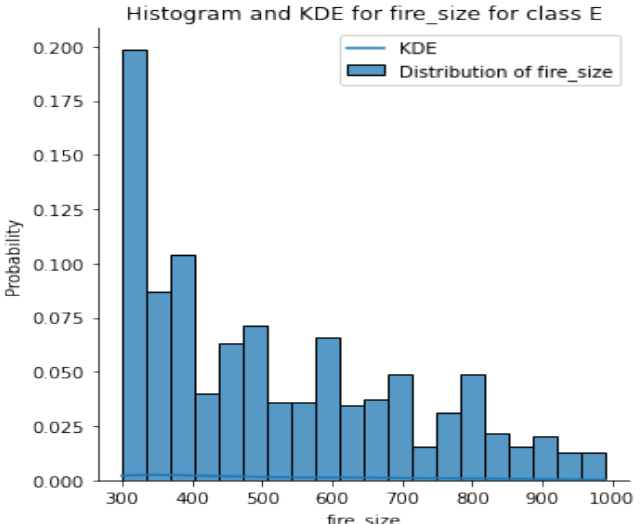
1.1 Modelling fire_size from the perspective of the fire_size class:

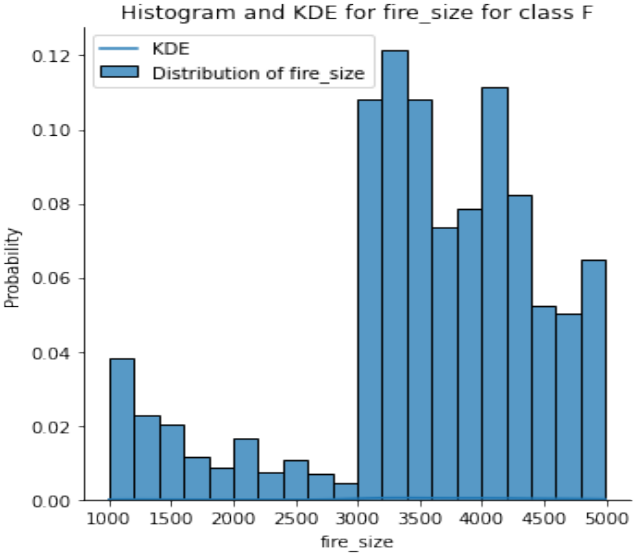
For fire_size we decided to take only one class as an example — class B. This is the class where the most data is located and the probabilities are highest. For other classes we decided to show just how KDE and histograms are distributed..

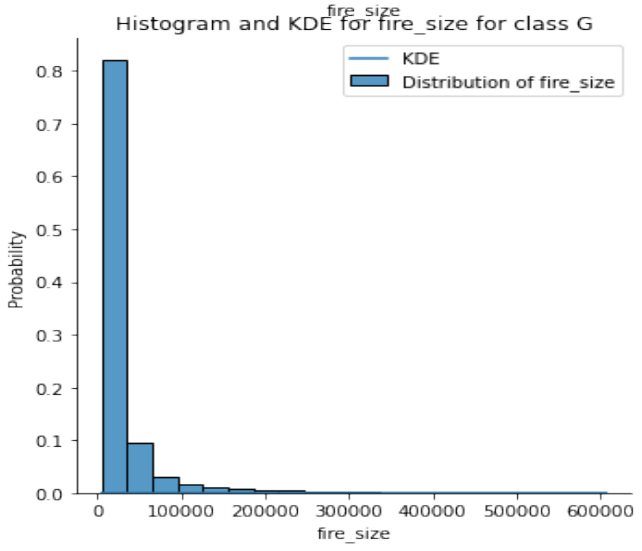


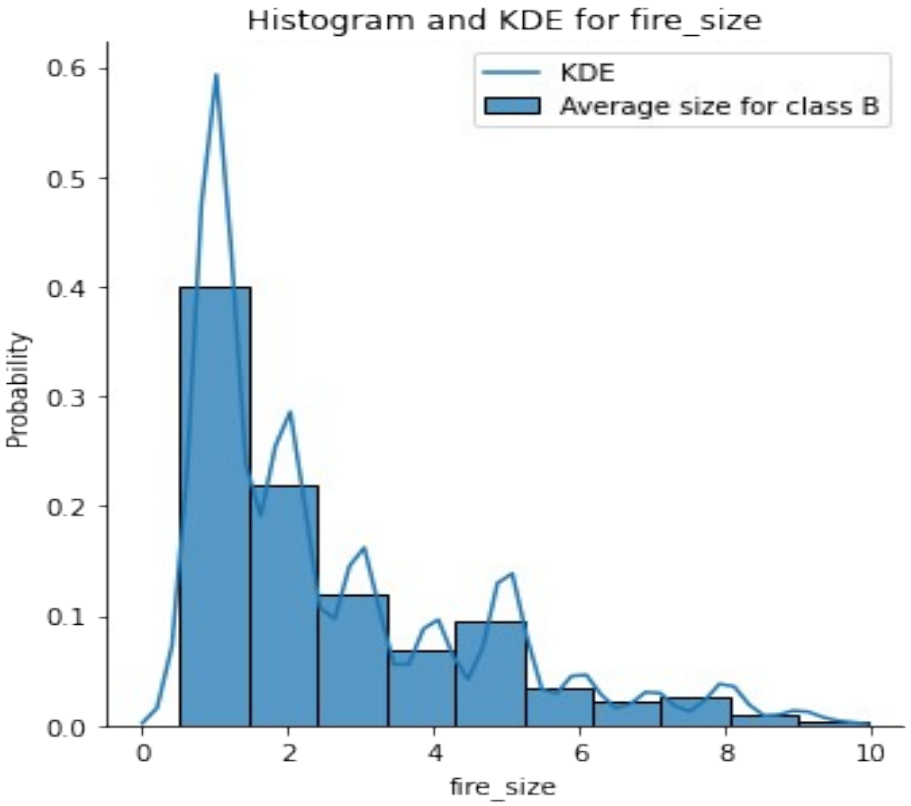


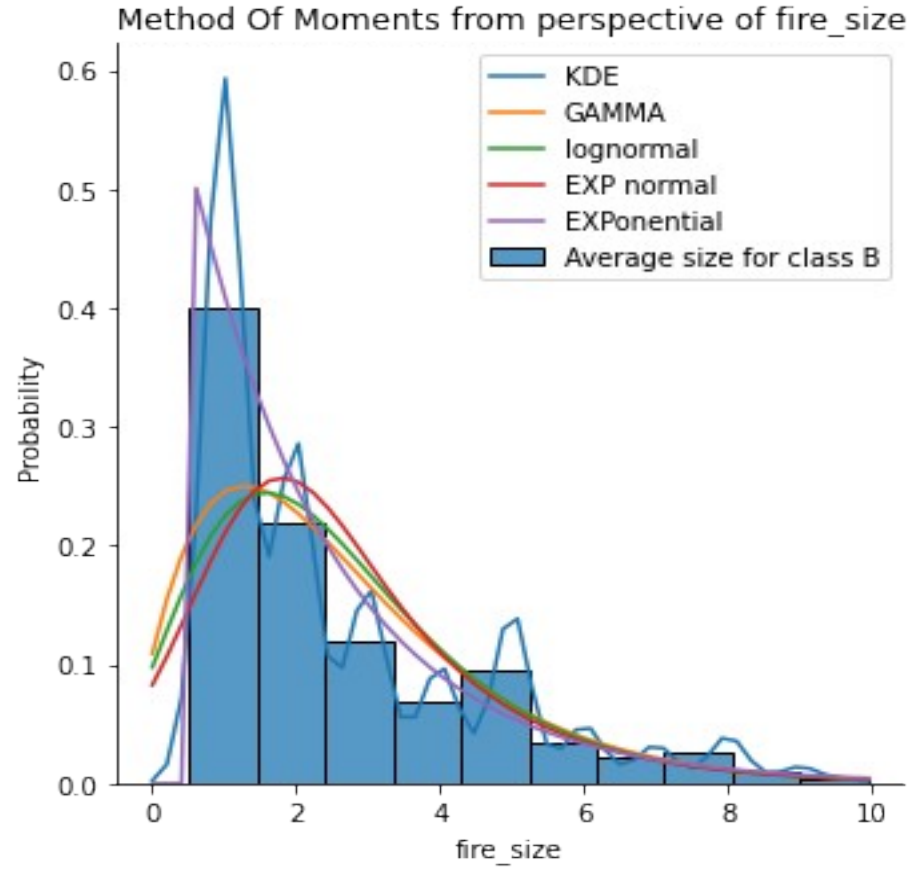




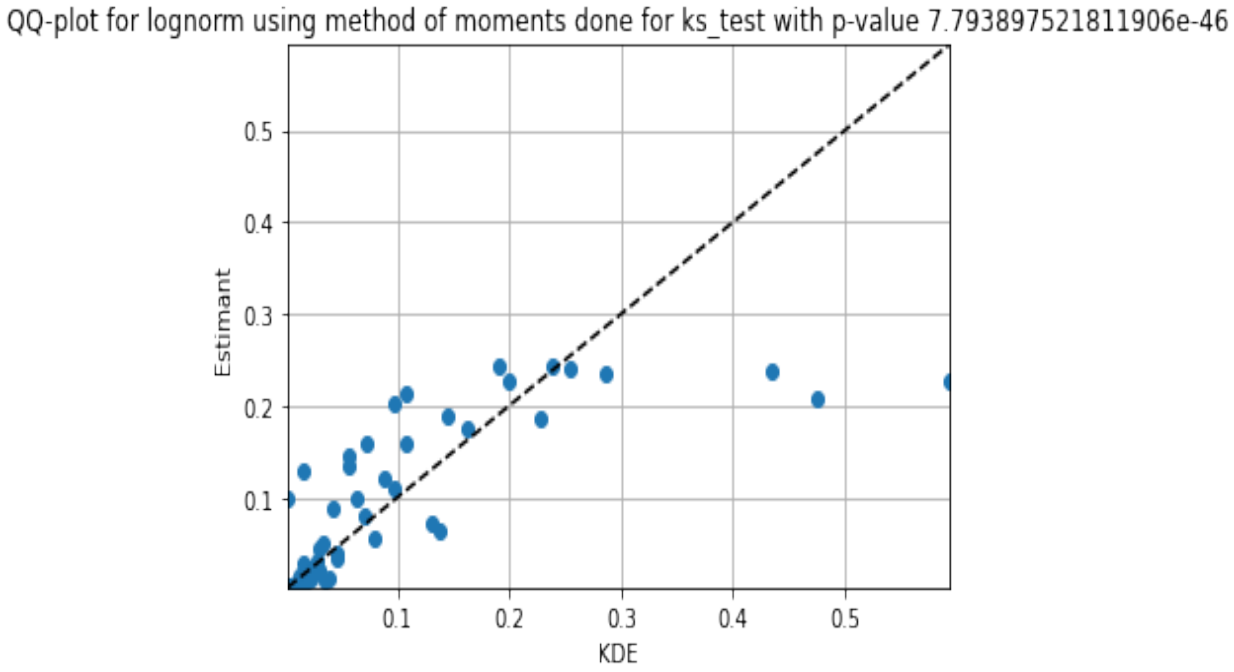




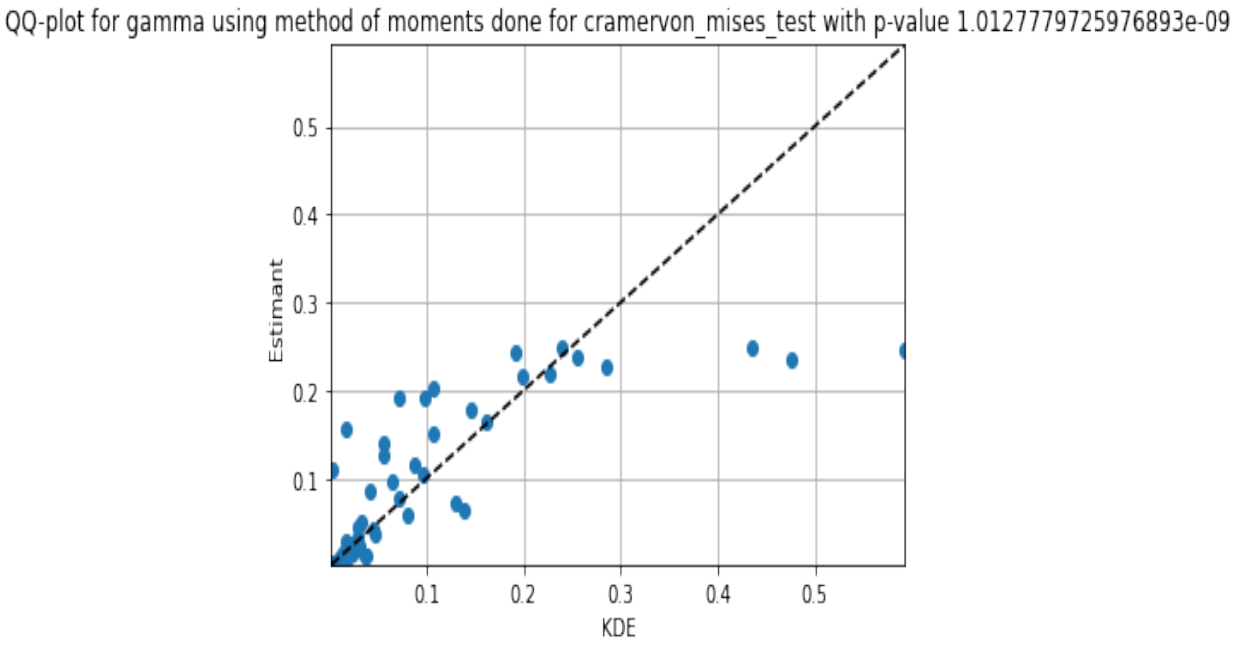

Histogram and KDE for fire_size
KDE
Average size for class B
Probability
fire_size
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8
10

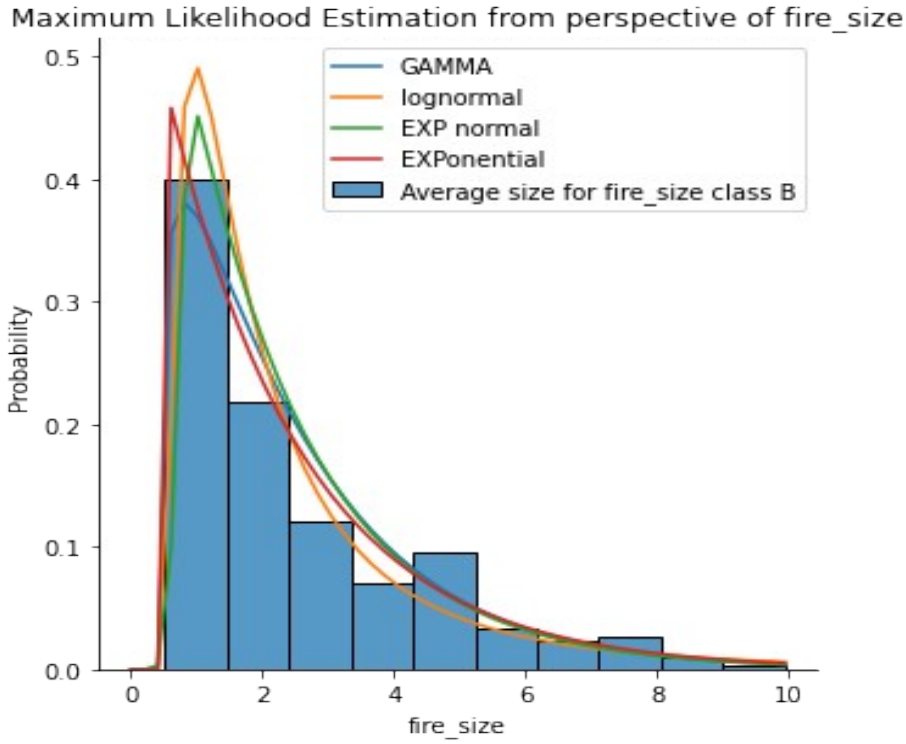

Method Of Moments from perspective of fire_size
KDE
GAMMA
lognormal
EXP normal
EXPonential
Average size for class B
Probability
fire_size
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8
10

Best fit according to Kolmogorov-Smirnov test:

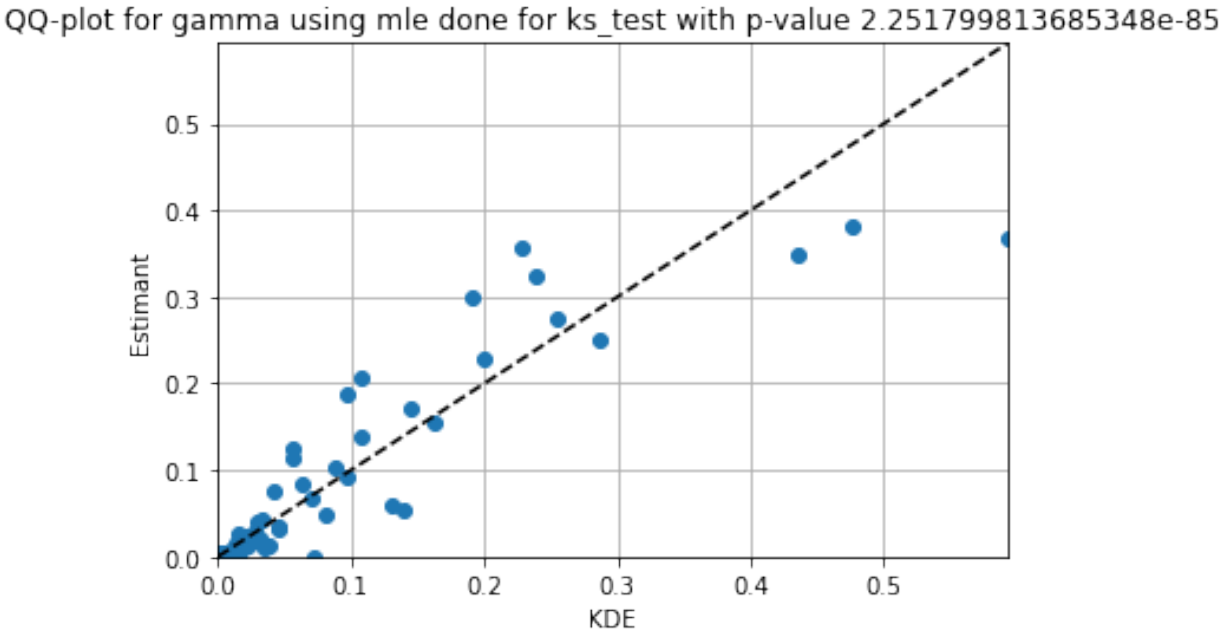


Best fit according to Cramer-von Mises test:

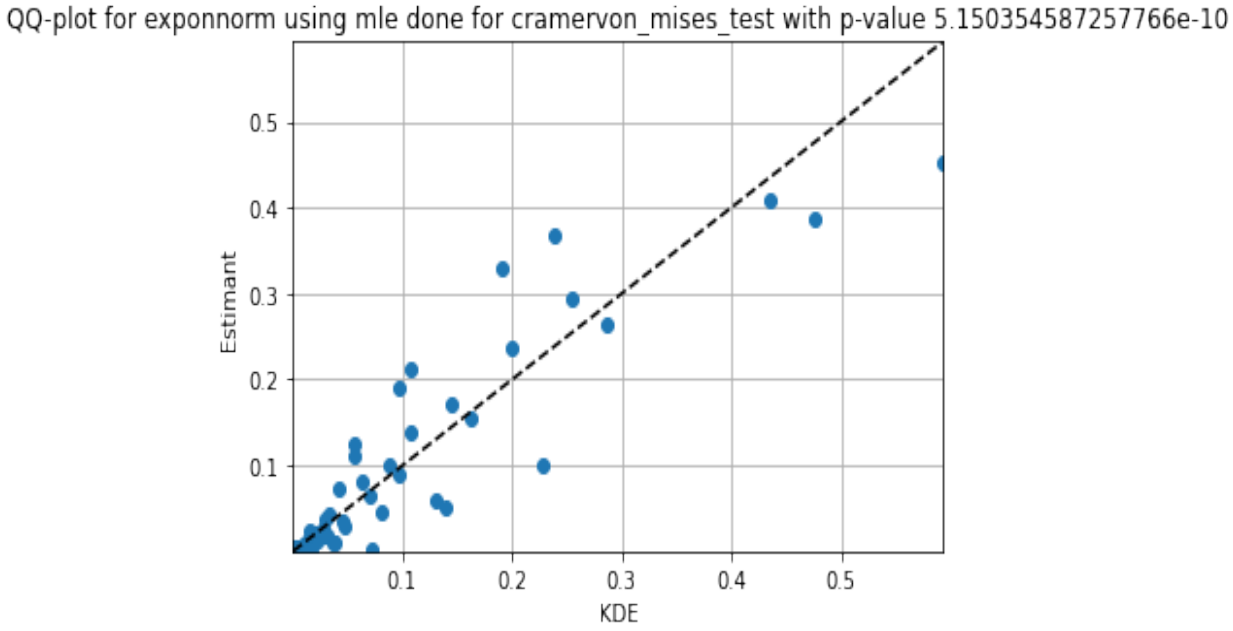


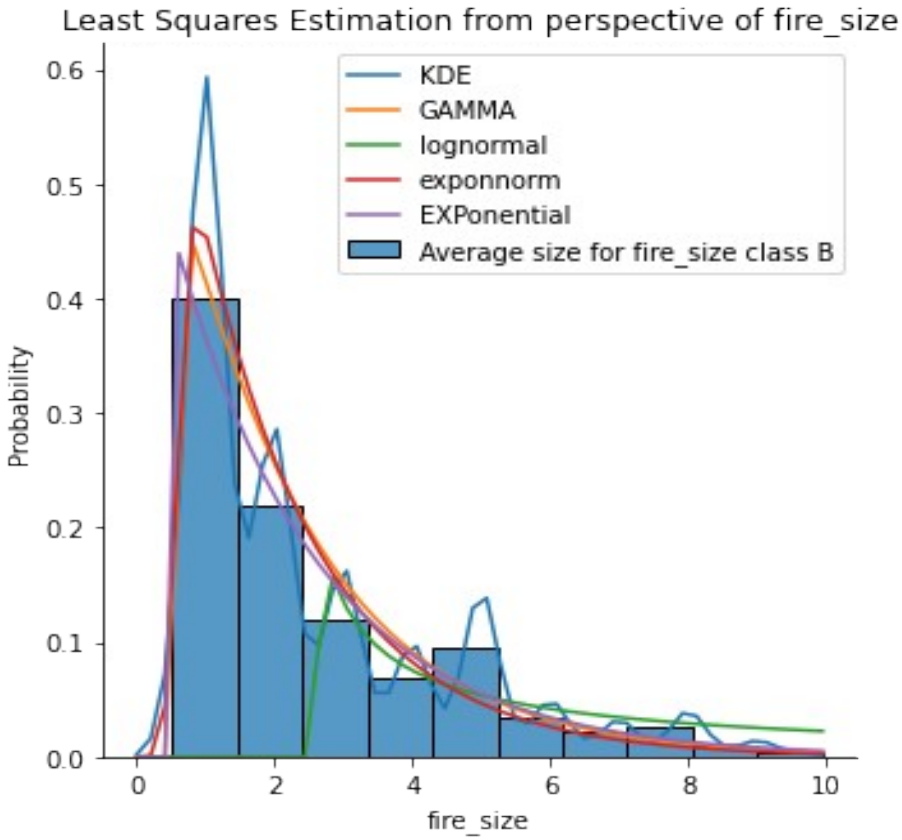


Best fit according to Kolmogorov-Smirnov test:

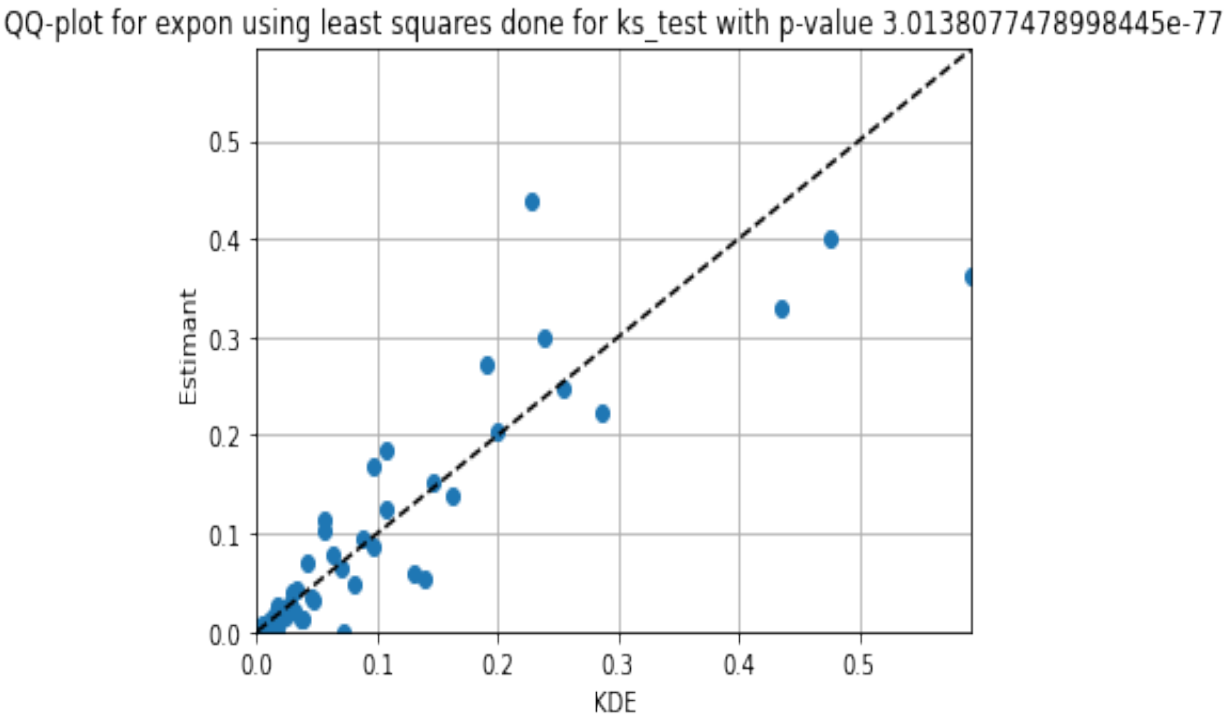


Best fit according to Cramer-von Mises test:

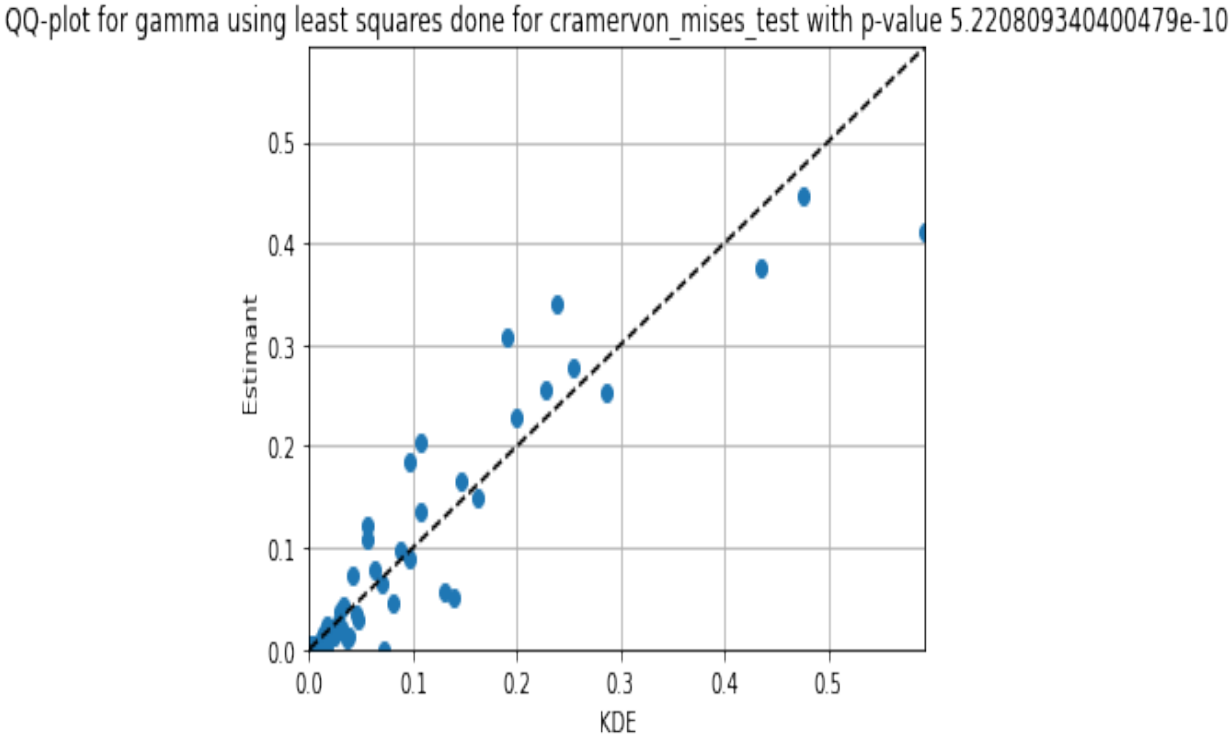




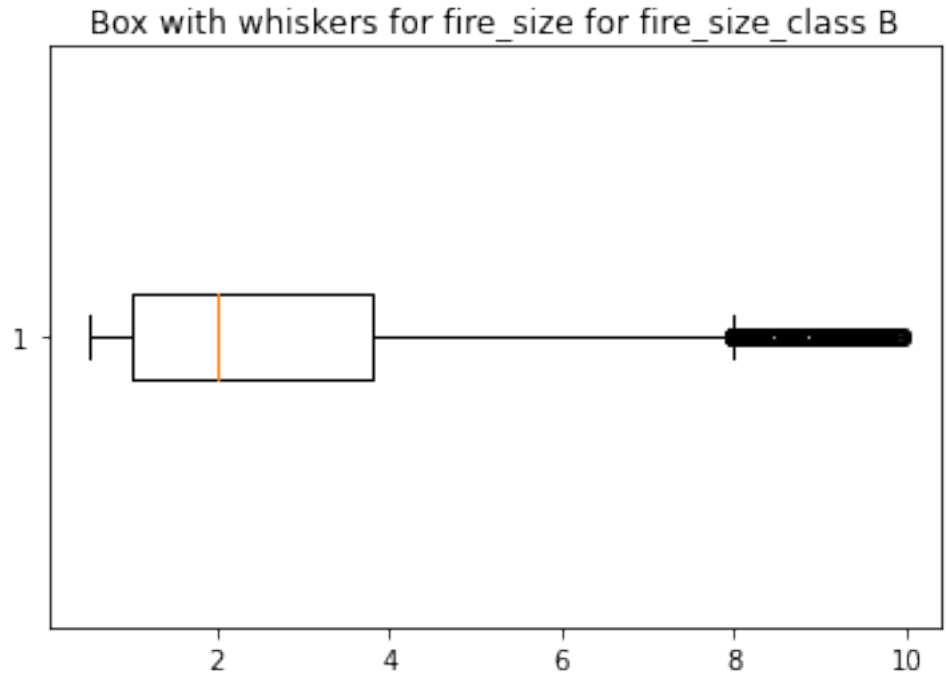
Best fir for Kolmogorov-Smirnov test:



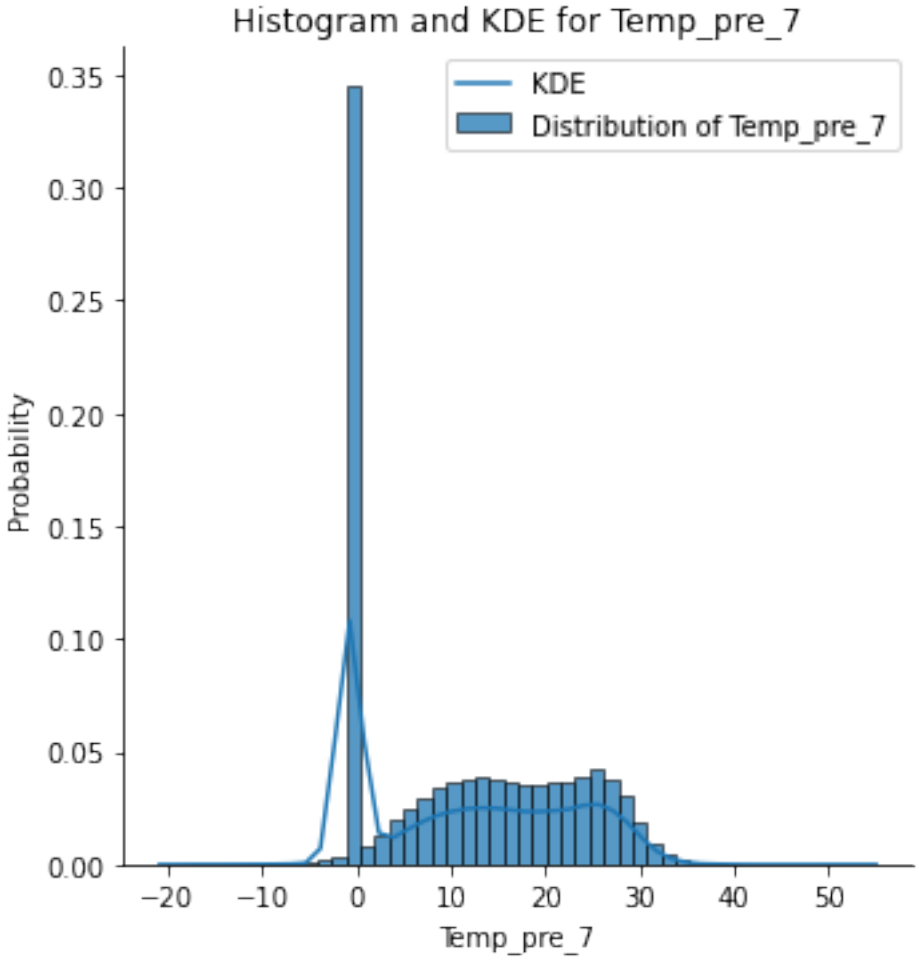
Best fit according to Cramer von Mises test:

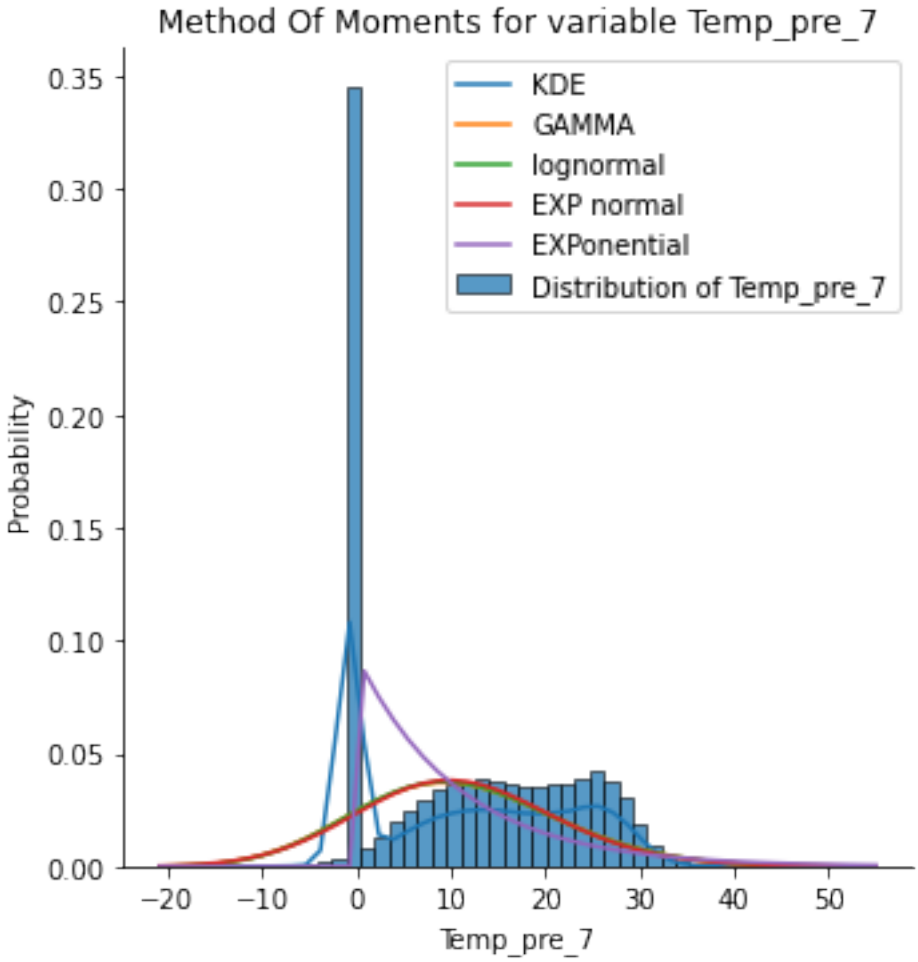


Box with whiskers for fire_size:

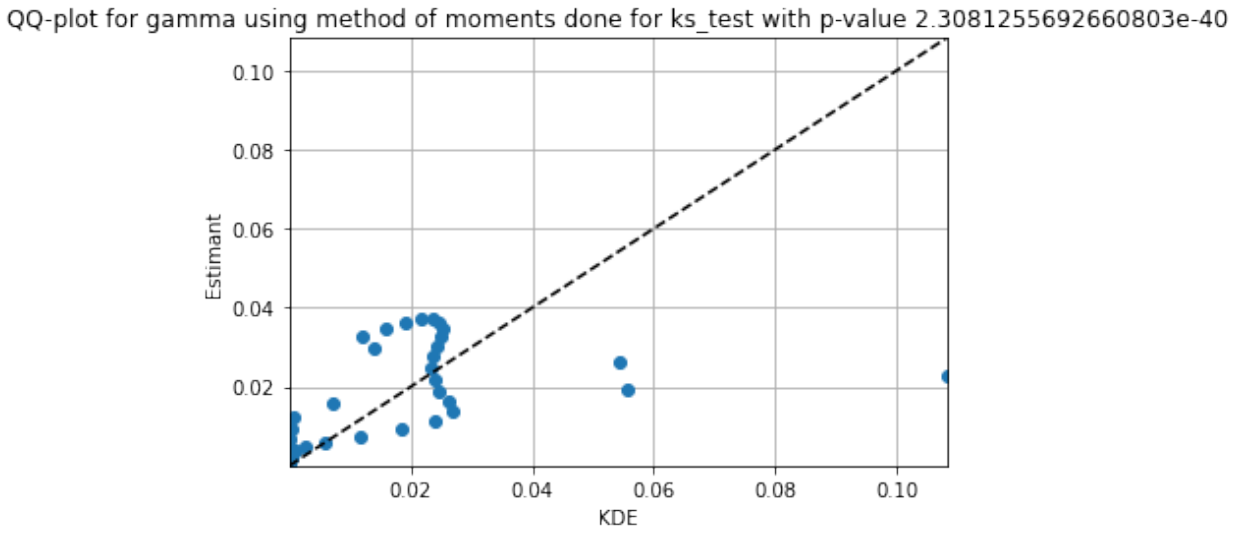


1.2 Modelling Temp_pre_7:

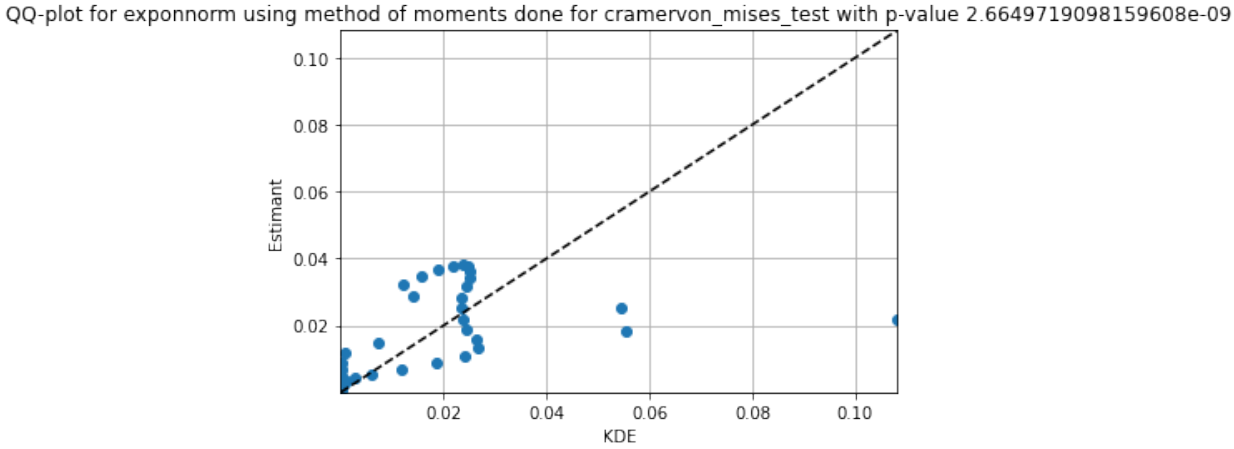


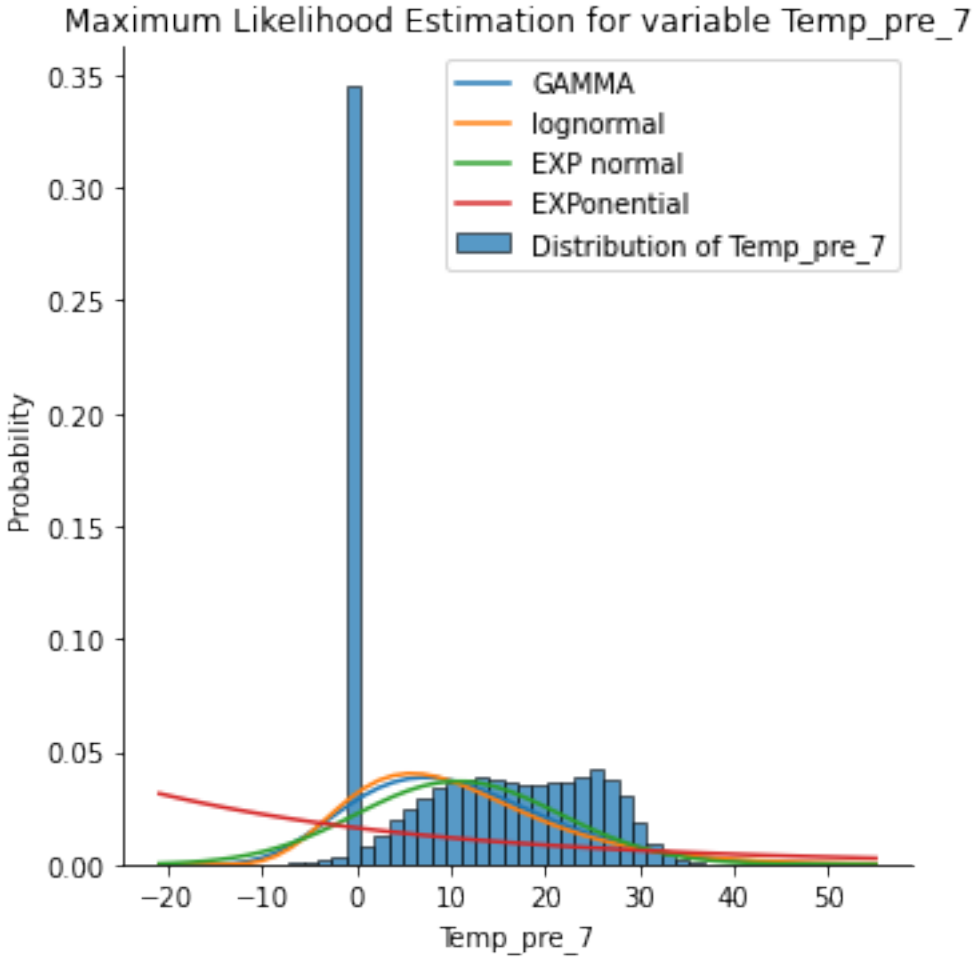


Best fit accoridng to Kolmogorov Smirnov test

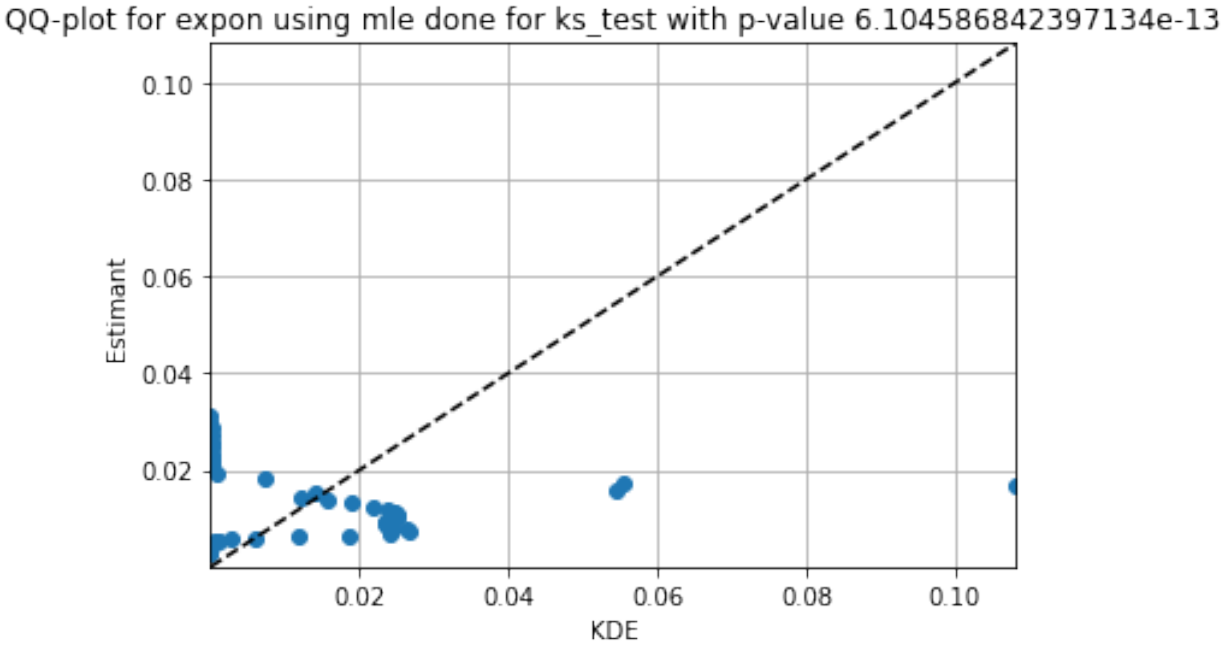


Best fit according to Cramer-von Mises test:

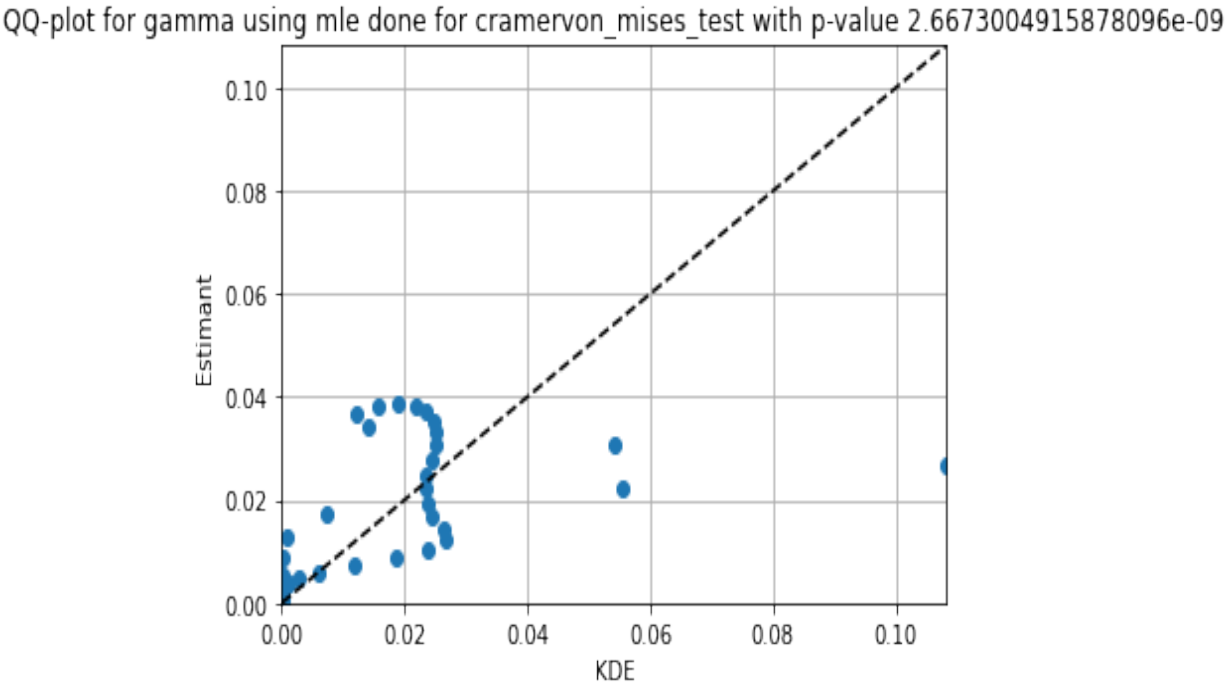


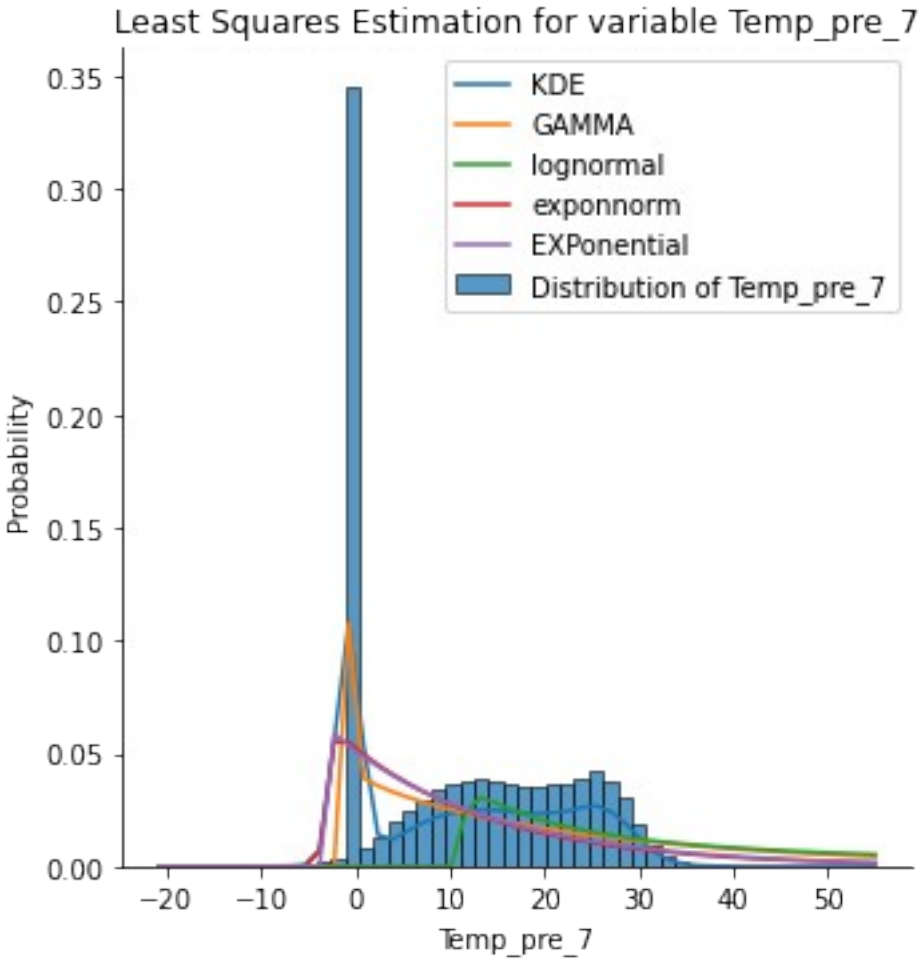


Best fit accoridng to Kolmogorov Smirnov test

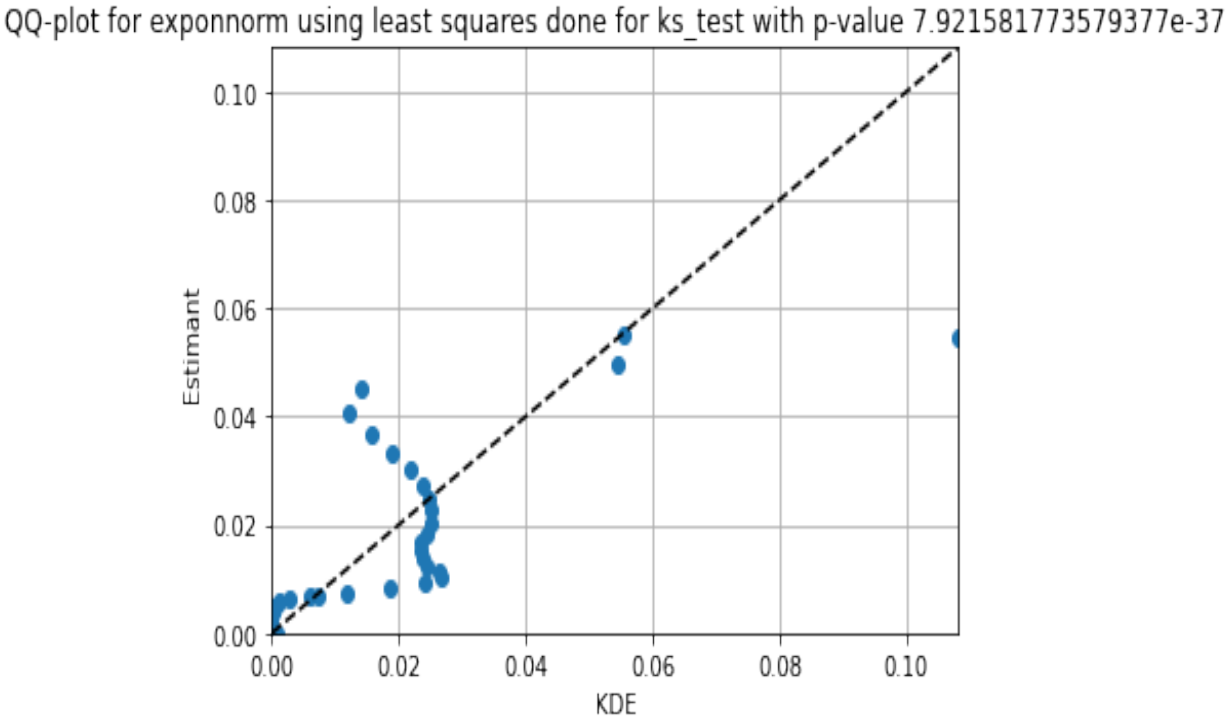


Best fit according to Cramer-von Mises test:

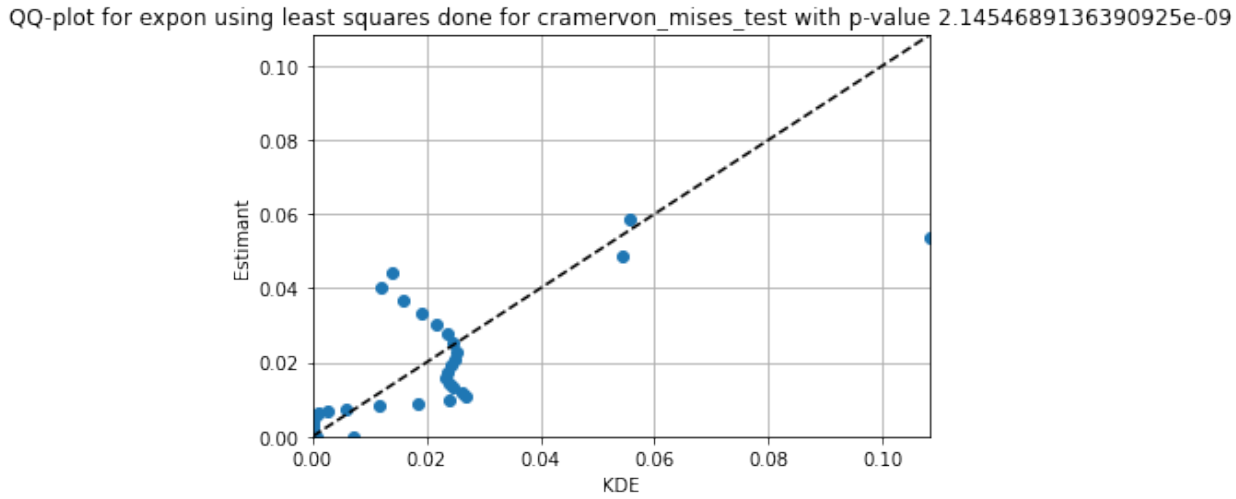




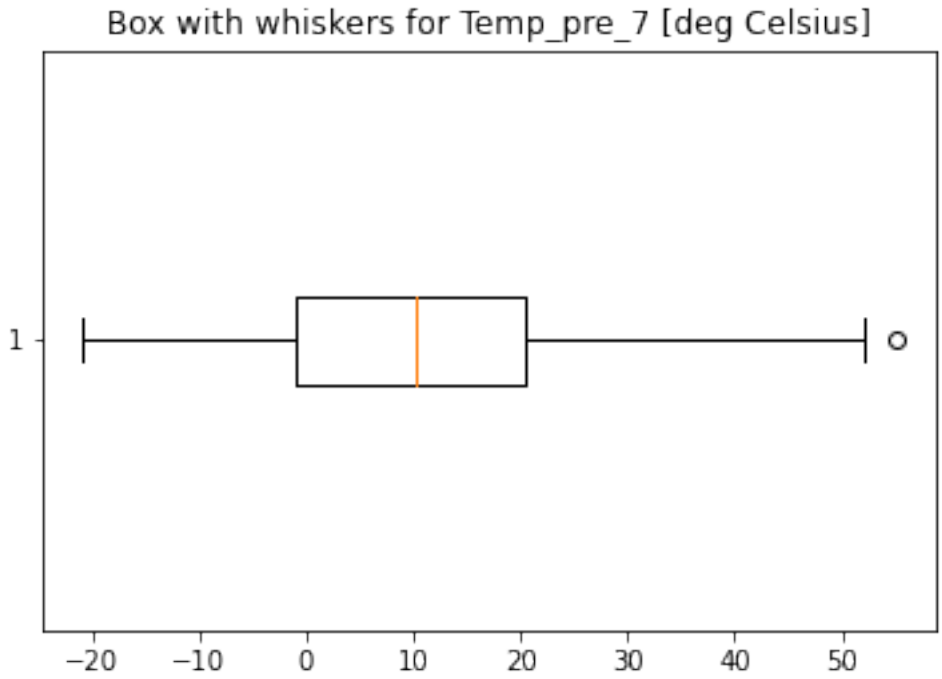
Best fit accoridng to Kolmogorov Smirnov test



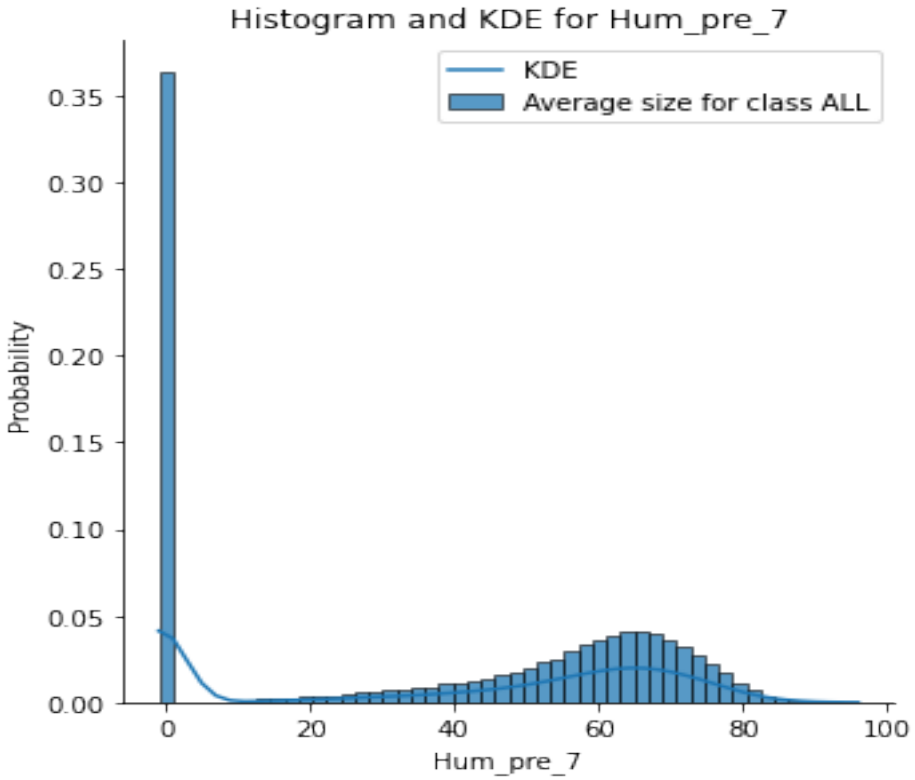
Best fit according to Cramer-von Mises test:

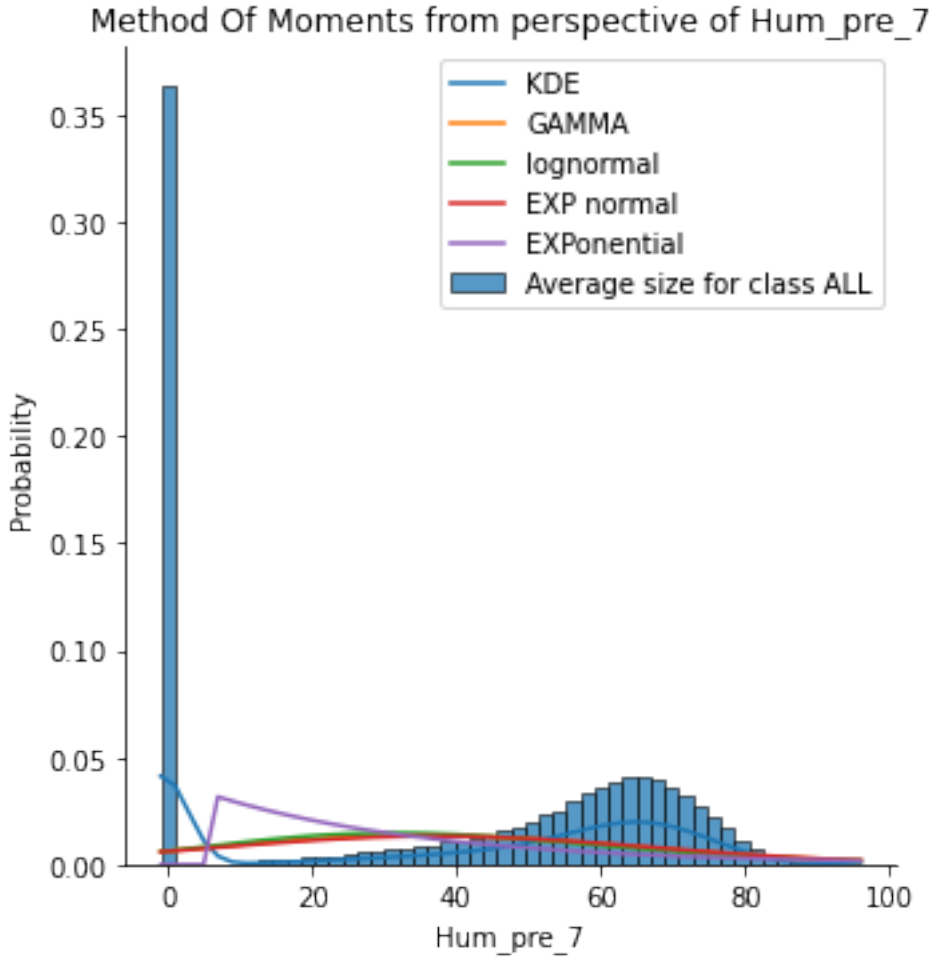


Box with whiskers for Temp_pre_7

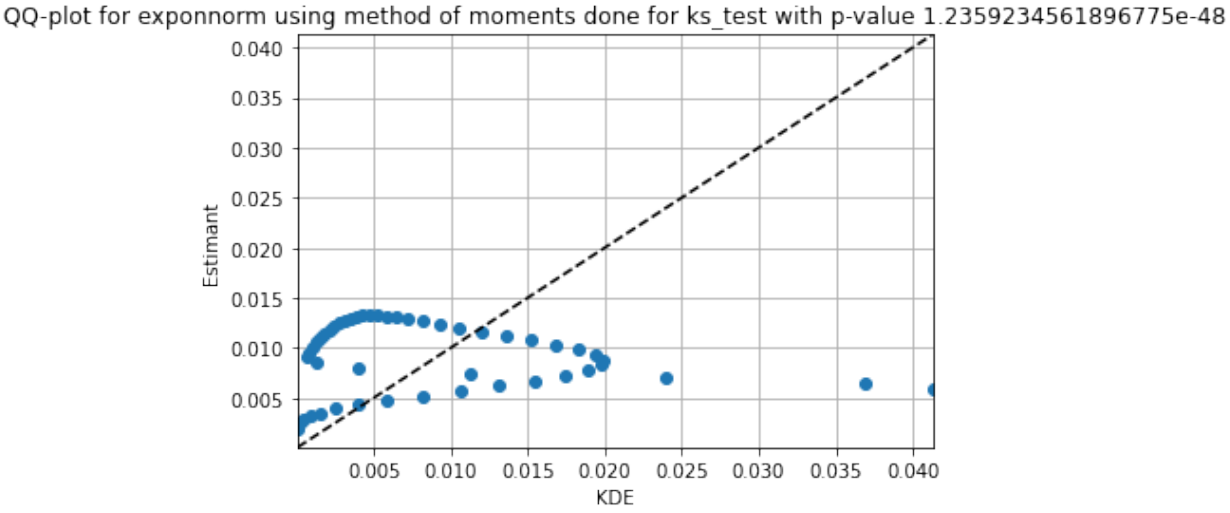


1.3 Modelling Hum_pre_7

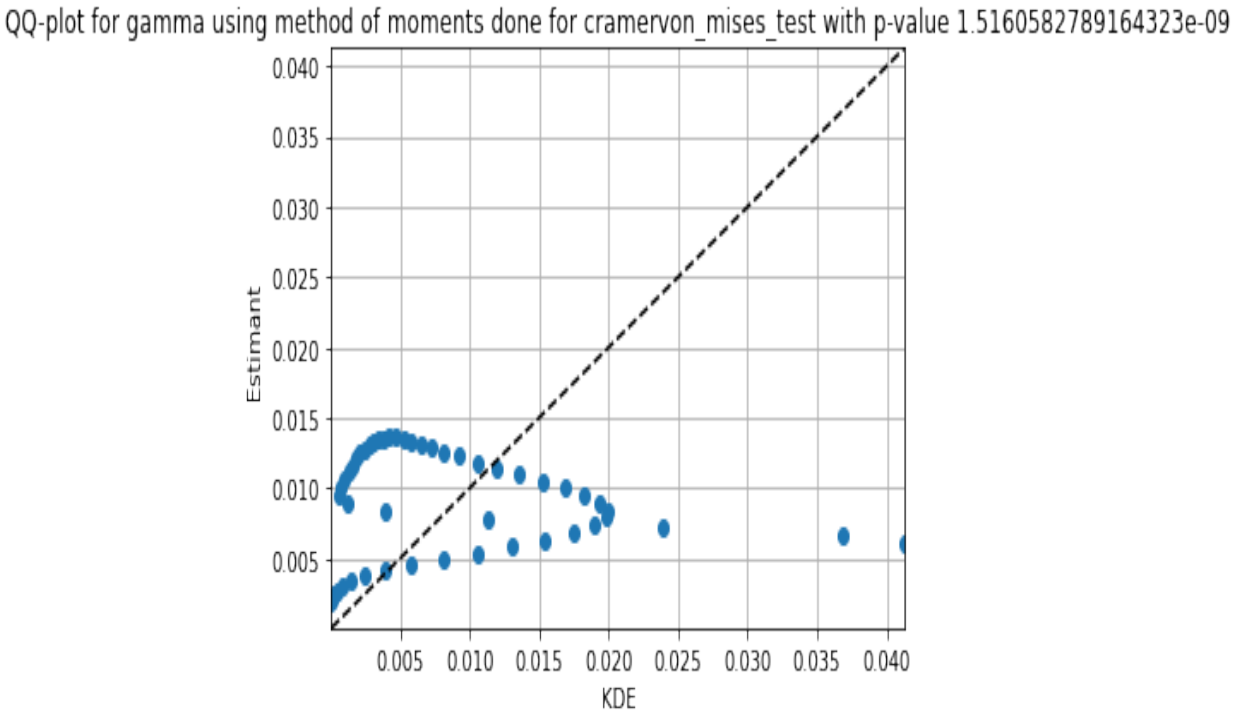


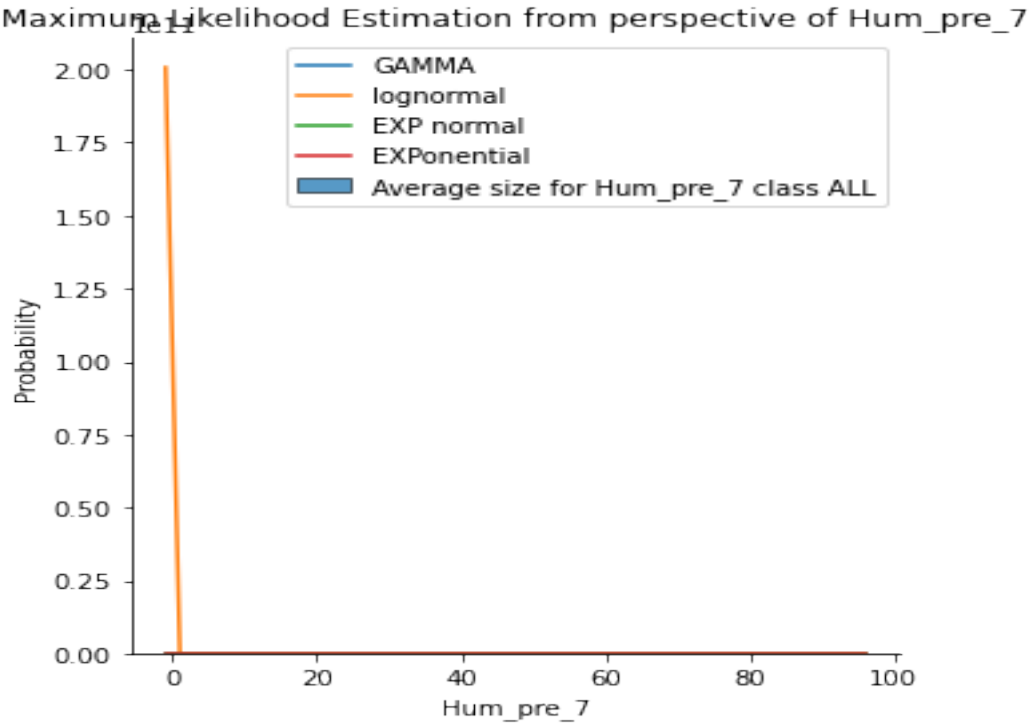


Best fit according to Kolmogorov-Smirnov test:

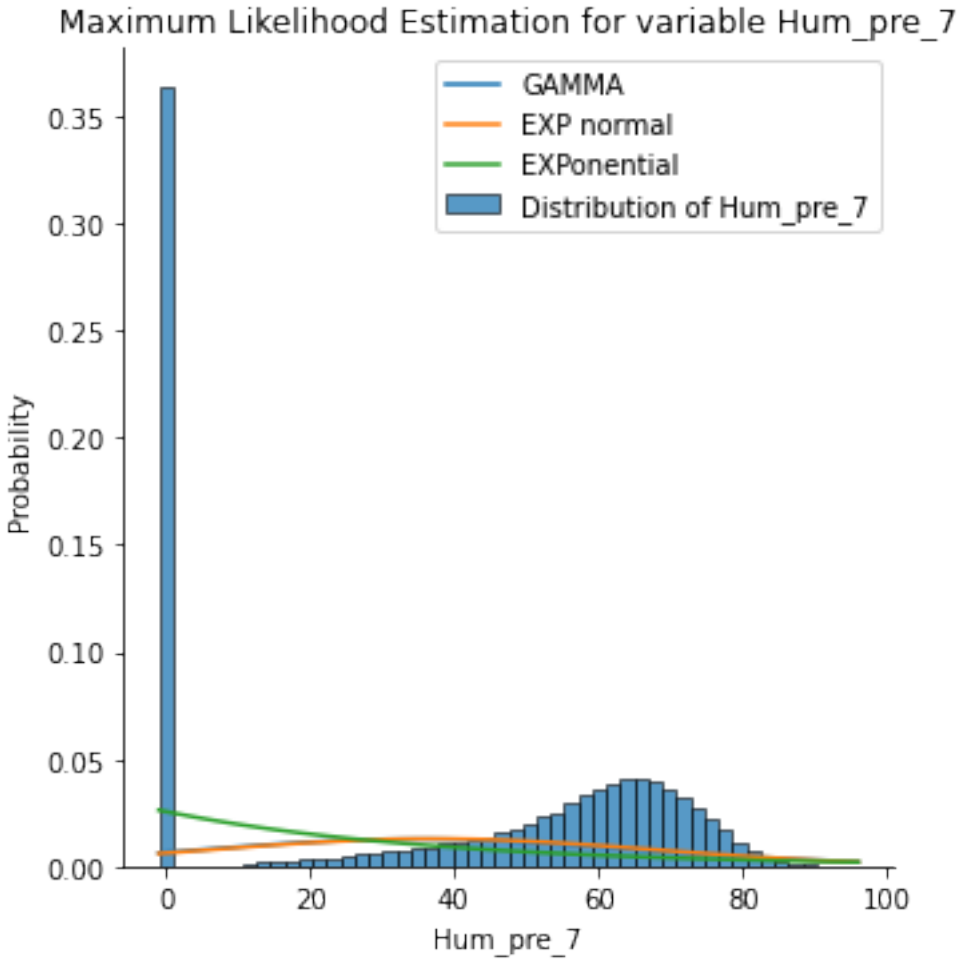


Best fit according to Cramer-von Mises:

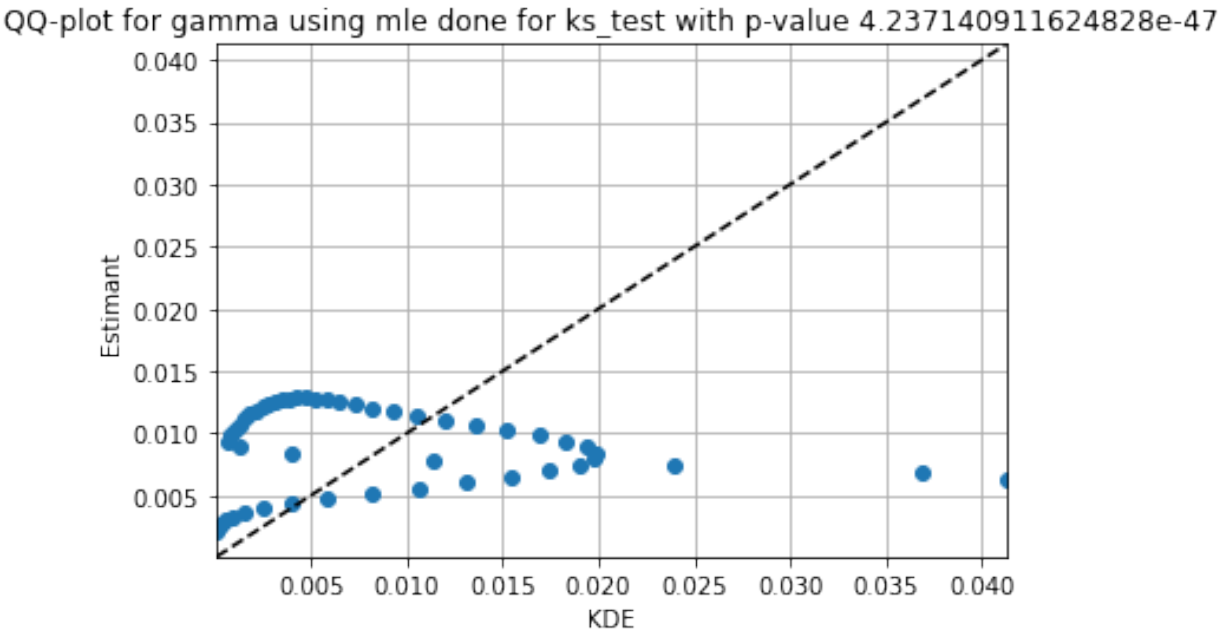




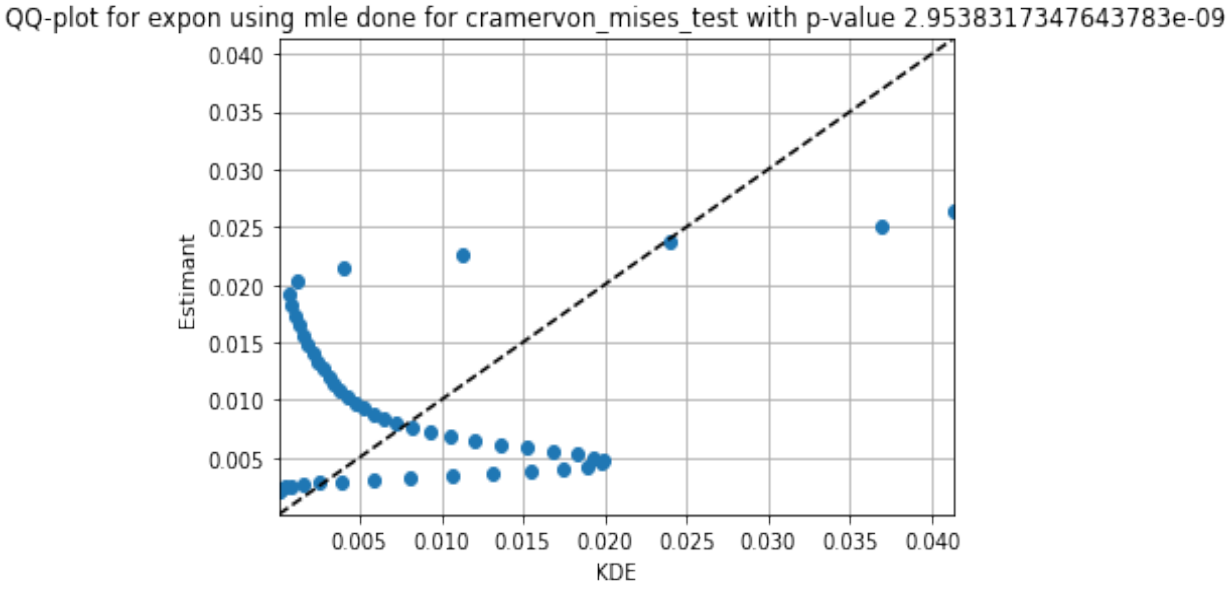
As we can see the maximal likelihood estimation predicted the distribution quite unnaturally in this case. Based on that in this case excluding lognormal is the best option.

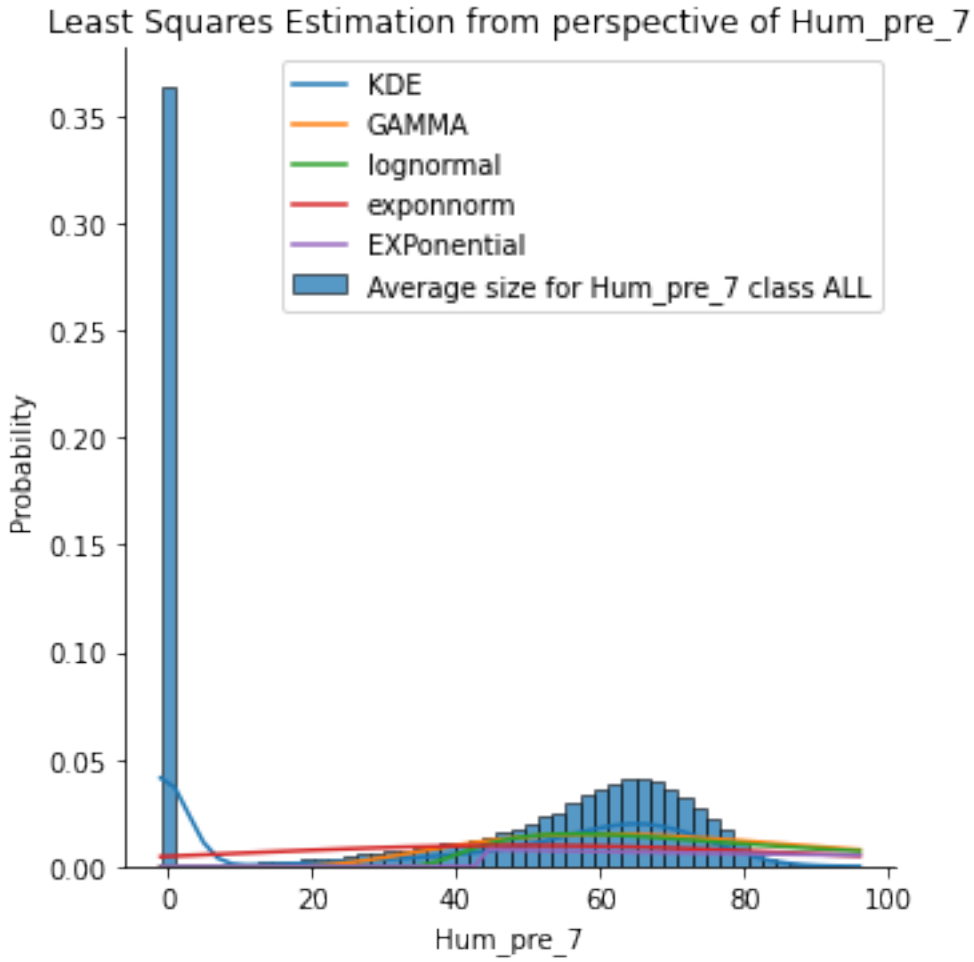


Best fit according to Kolmogorov-Smirnov test:

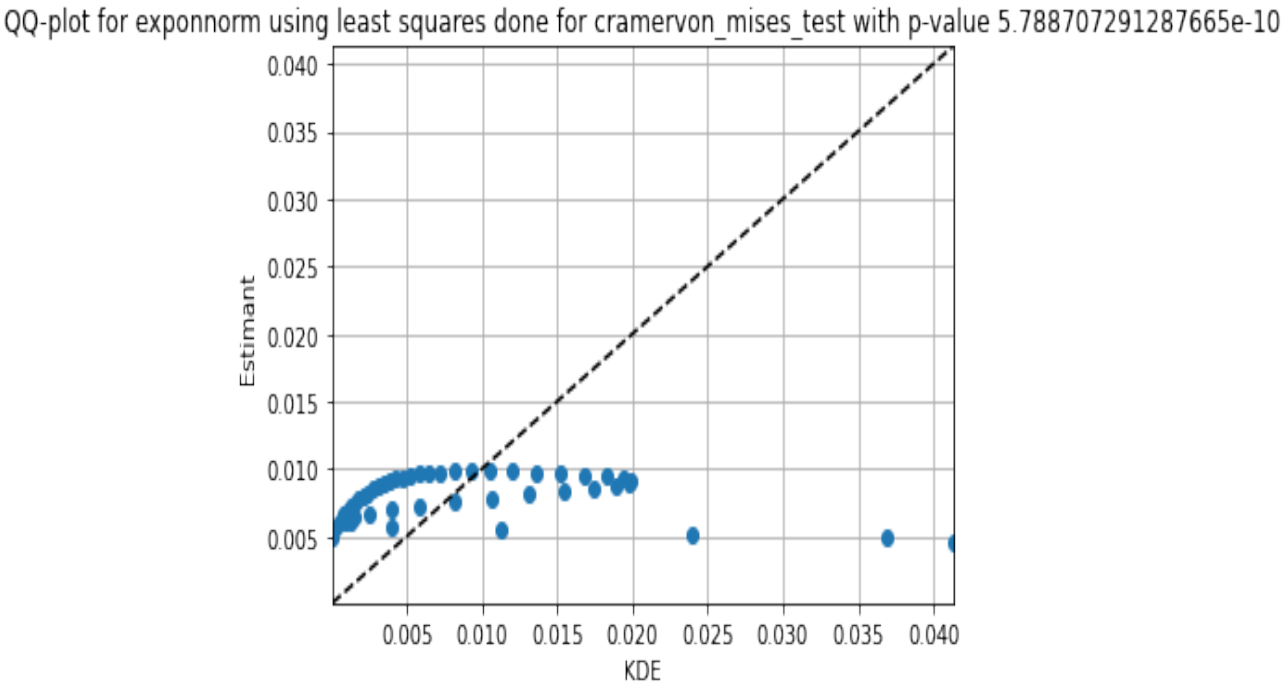


Best fit according to Cramer-von Mises test

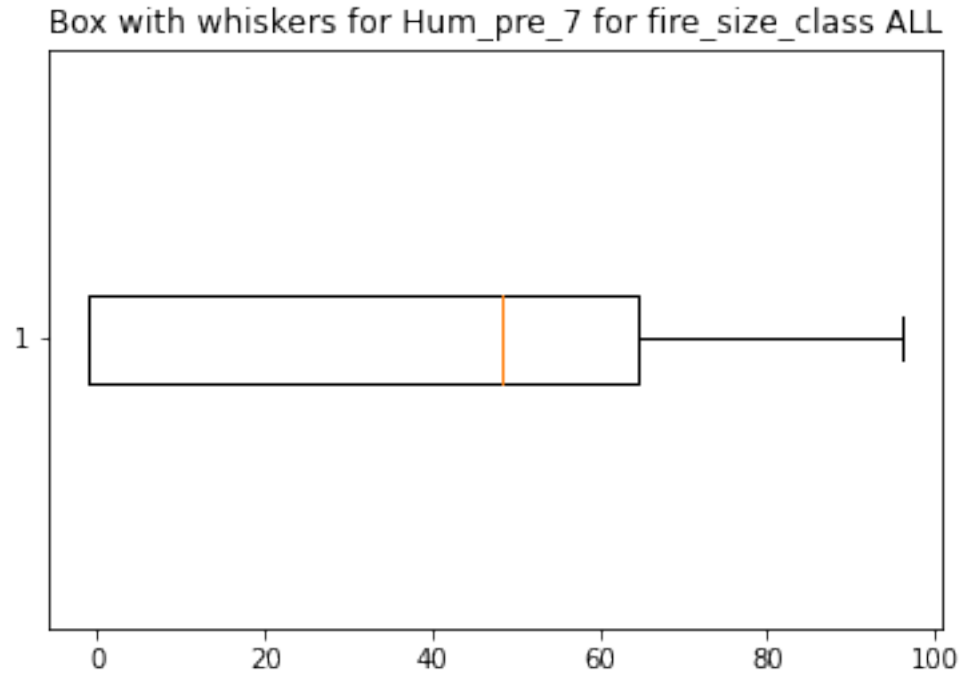




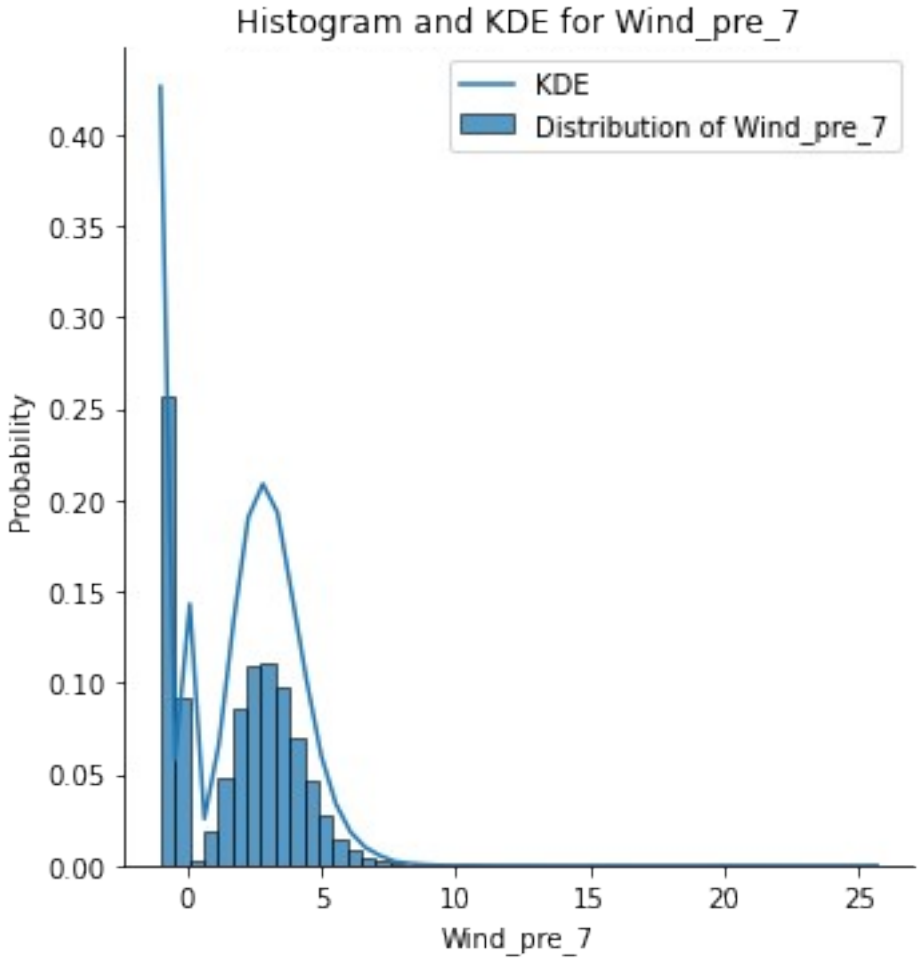
Best fit according to both Kolmogorov-Smirnov(statistic=0.8857657193444326, pvalue=1.60024018490612e-47) and Cramer-Von Mises(statistic=11.612295710969551, pvalue=5.788707291287665e-10) tests

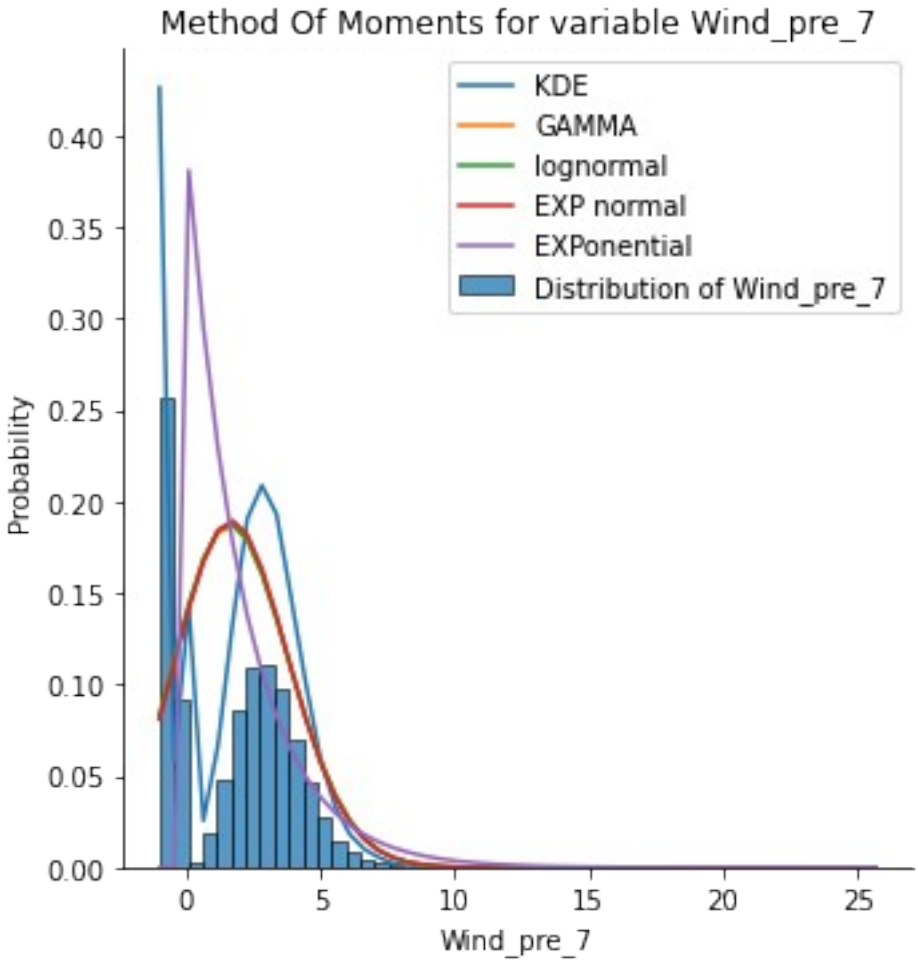


Box with whiskers for this variable is:

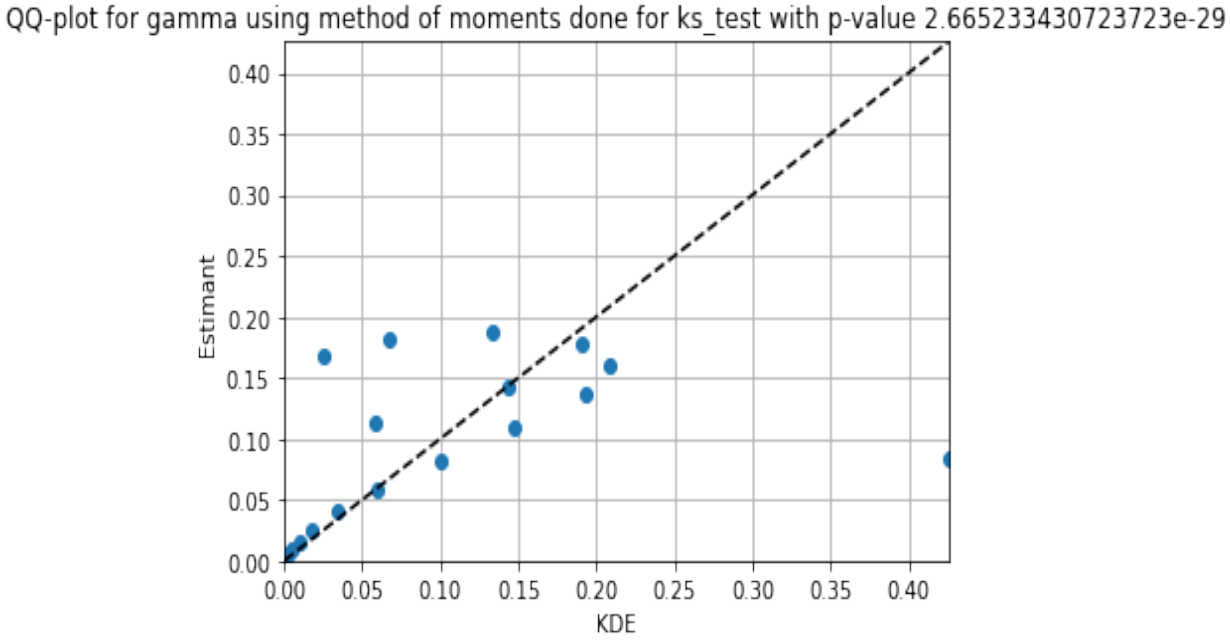


1.4 Wind_pre_7

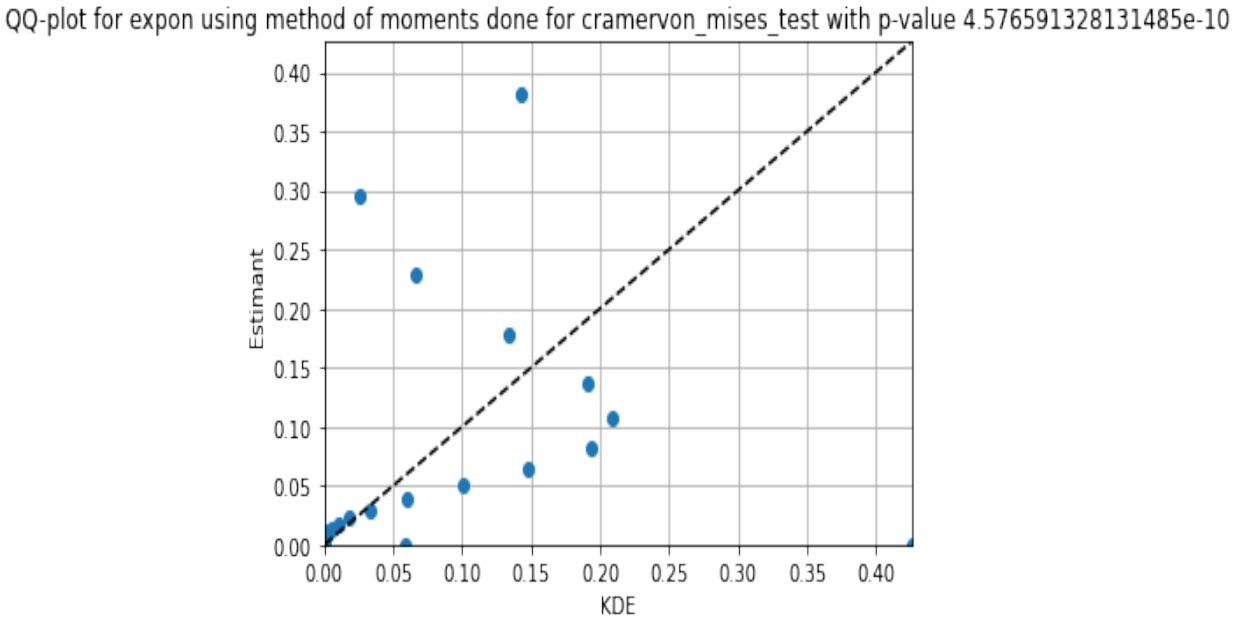


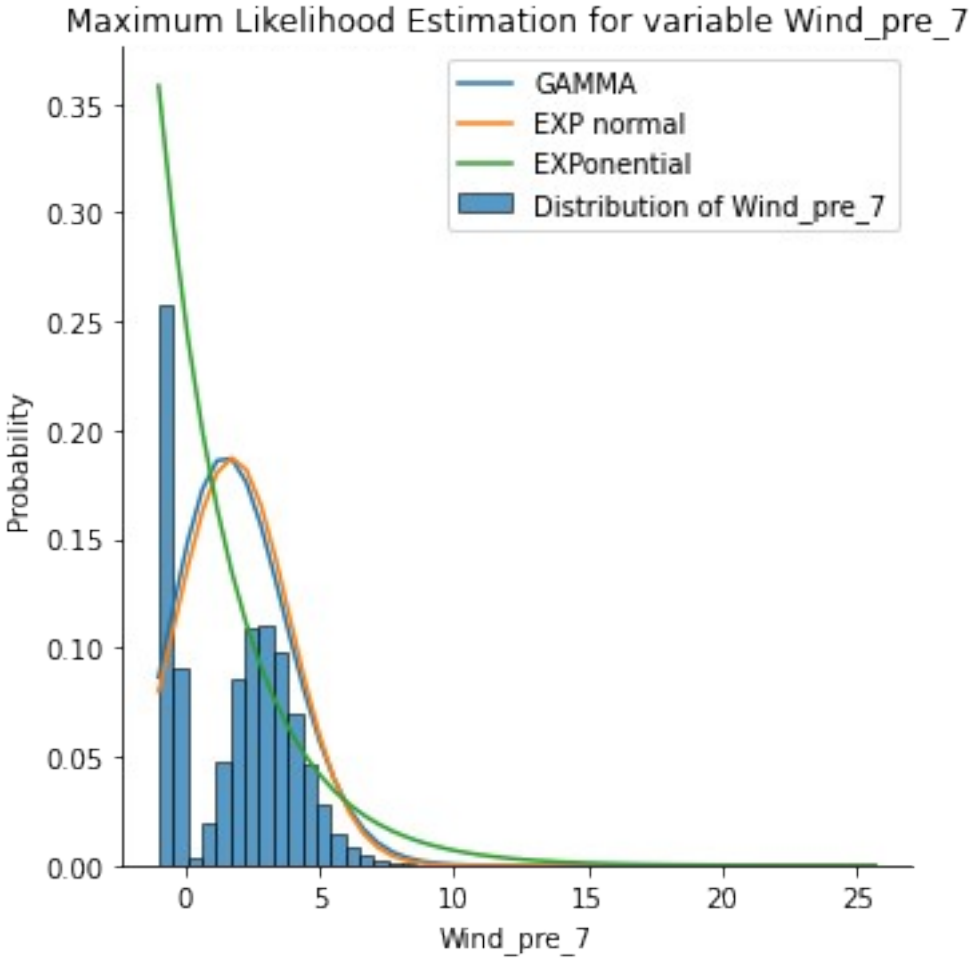


Best result for KS:



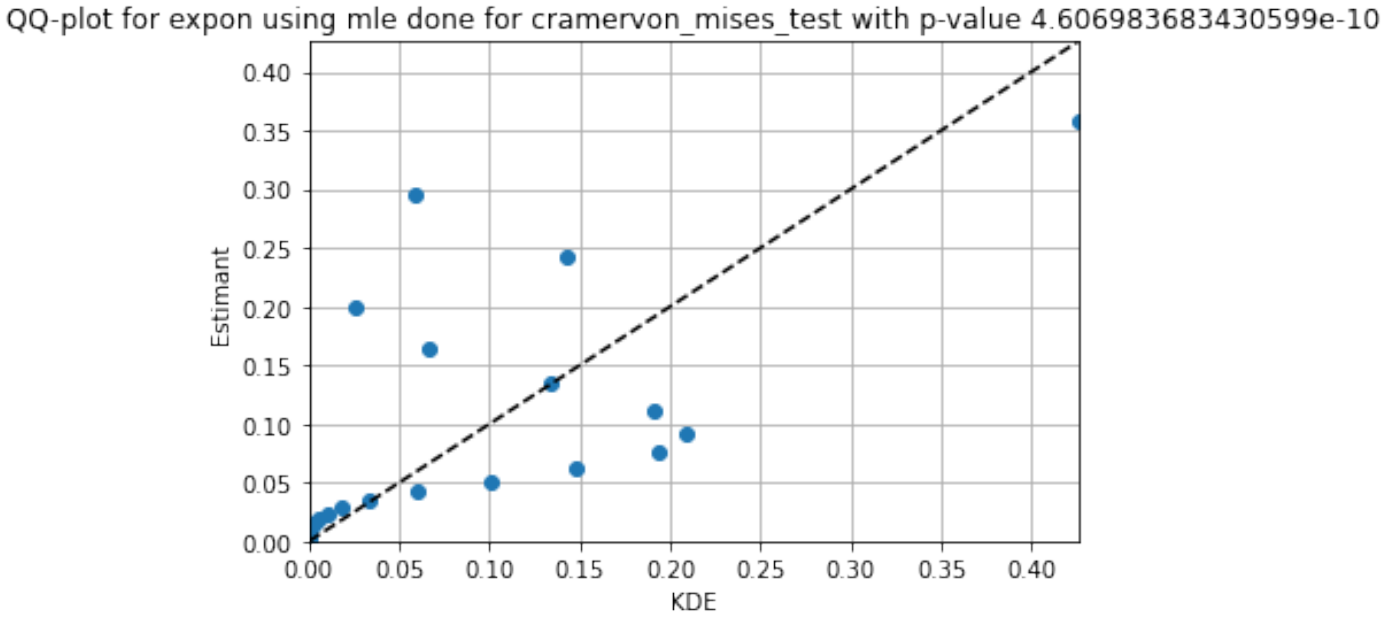
Best result for Cramer-von Mises:

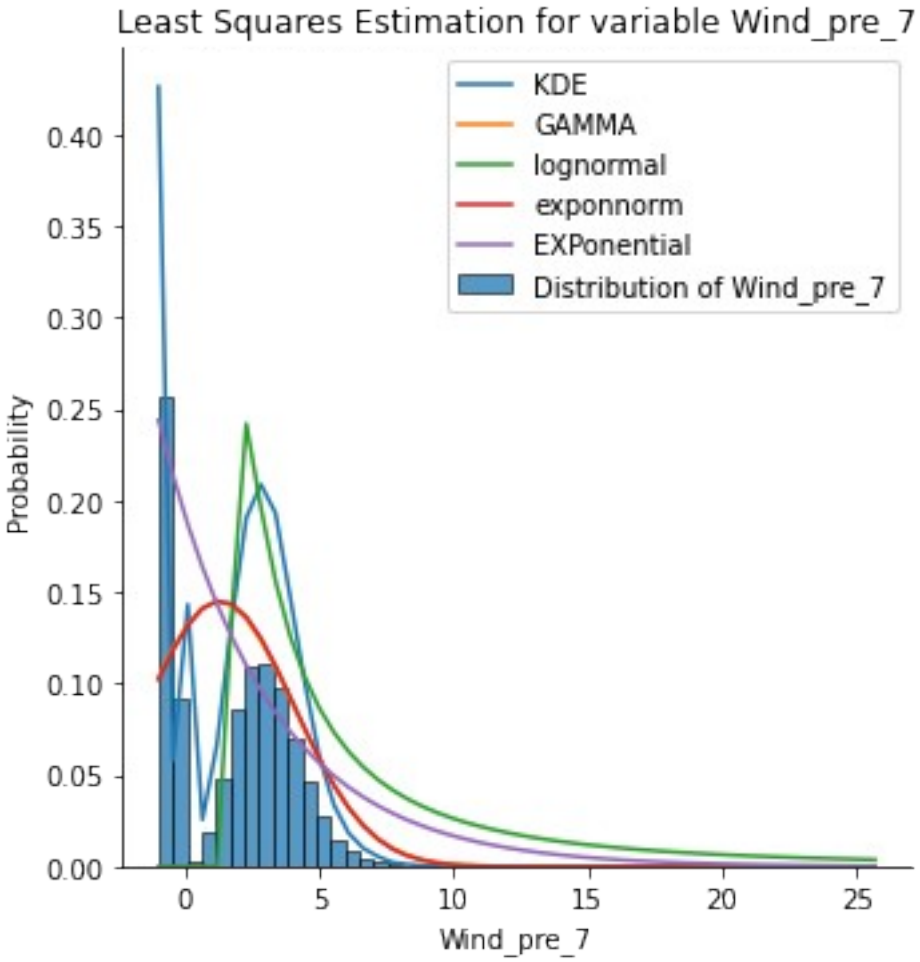




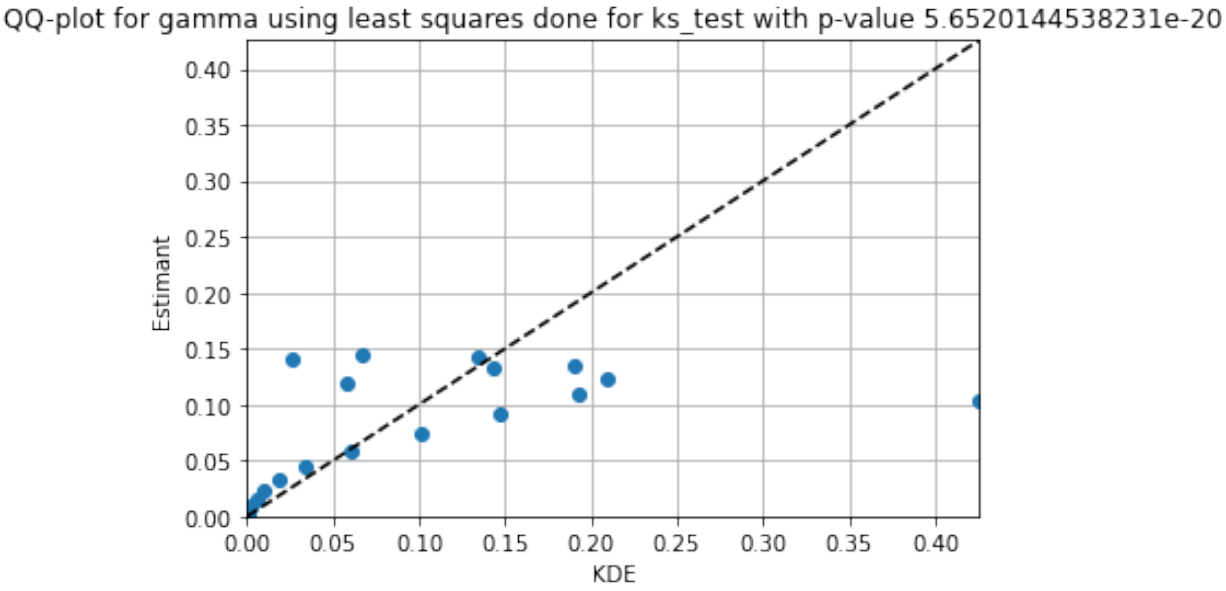
Similarly as in the previous case, lognormal was not modelled correctly, therefore it was excluded from here.

According to both tests exponential function modelled the phenomena best. For KS-test(statistic=0.628541534279319, pvalue=1.2127339260734556e-19), for Cramer von Mises(statistic=5.626606956264935, pvalue=4.606983683430599e-10)

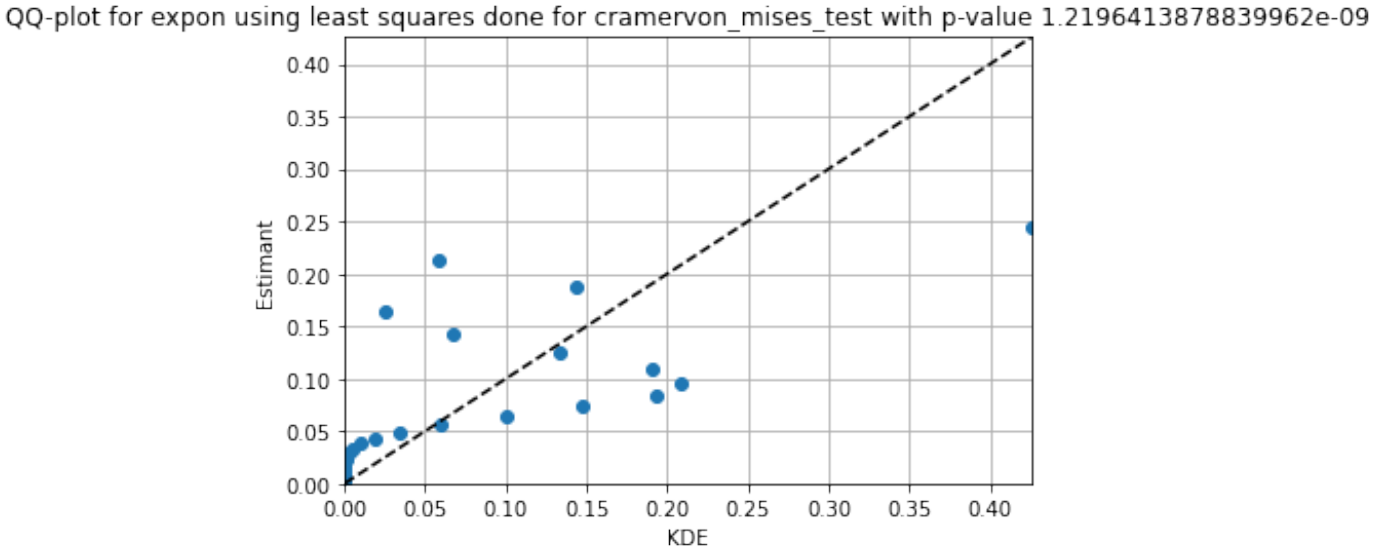




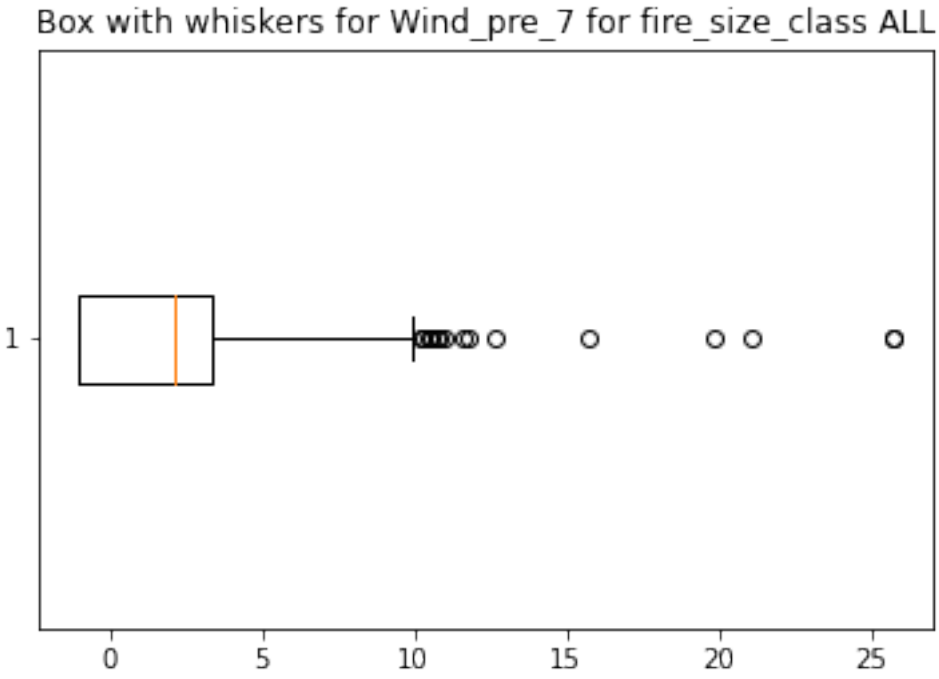
For Kolmogorov Smirnov the best result:



The best result for Cramer-Von Mises test:



Box with whiskers



Conclusion

As we can see, sometimes not only one distribution can be used to model some phenomena. Sometimes more than one can be applied and the difference between one and another is really small. Sometimes though, the distributions simply doo not apply and produce very bad results (as in case of lognorm for Hum_pre_7).

On top of that data in the real world observations can be distributed in many different ways. Based on that the distributions can help us understand the phenomena but usually cannot one-to-one reflect their random nature.

What is very useful in case of different methods are the tests. We used two of them — Kolmogorov-Smirnov (KS) and Cramer-von Mises. Both tests return two values: *statistic* and *p-value*. In assessment whether the theoretical distribution fits the real, p-value is the most useful. The bigger it is, the higher the probability the disribution fits the data. The data was however really sparsed and distributed sometimes in a very strange and ununiform matter — based on that very often the test results were rather poor and the probabilities really low (below 0.01). Visually, on the other hand, the distributions sometimes seemd to really well fit the data.

The tests were checked against the sample, not against the KDE. From this we can derive that simply the data in none of the cases is distributed using any of the suggested distributions, but the distributions still can serve as some kind of simplified models, especially that for some of the cases they really nicely fit to the histogram.

Sourcecode

https://github.com/PatrykStronski/MultivariateAnalysis_Task1

(branch *main*)