Enhanced report on learning practice # 1

Analysis of univariate random variables

Performed by:

Patrik Stronski

Ivan Pavelev

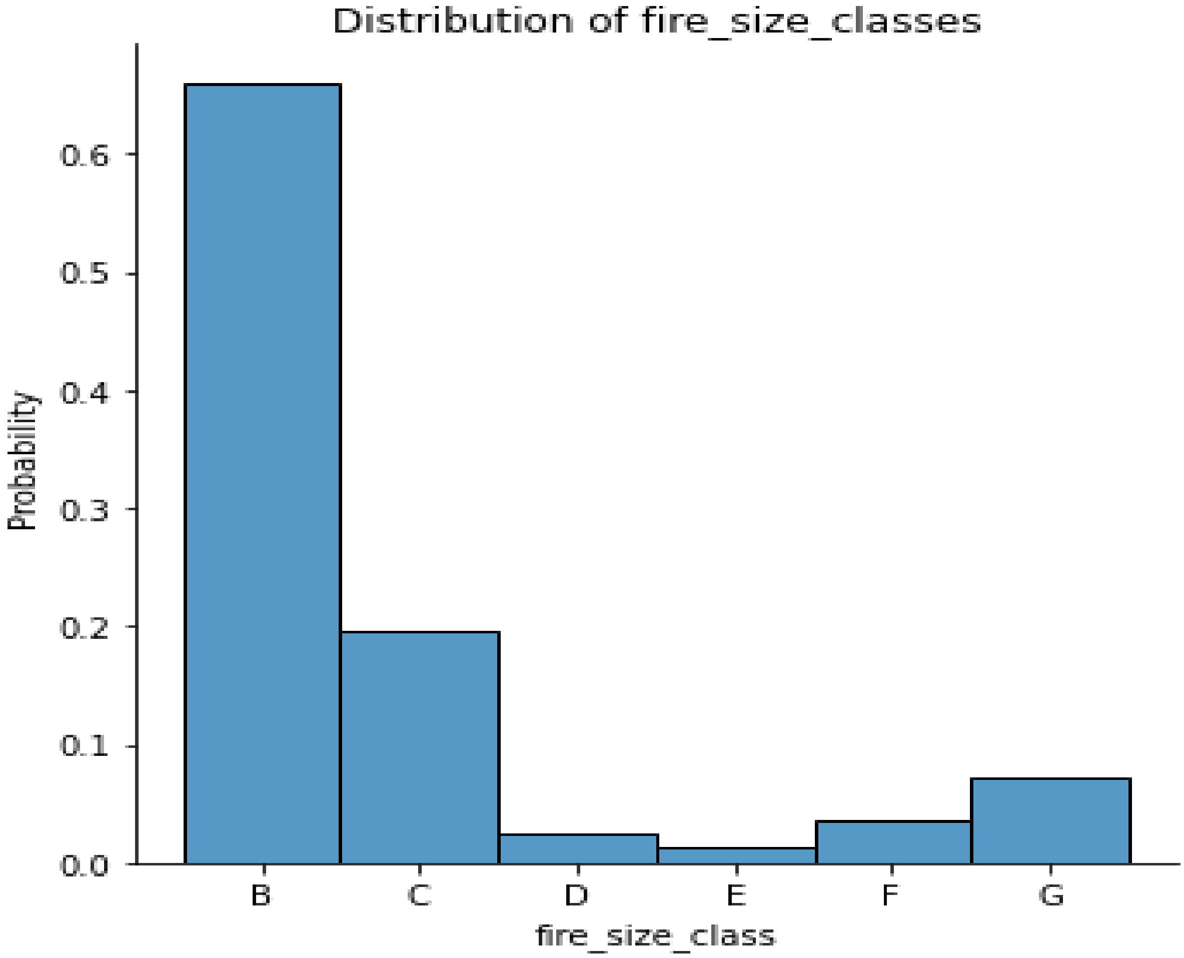
J4133C

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:

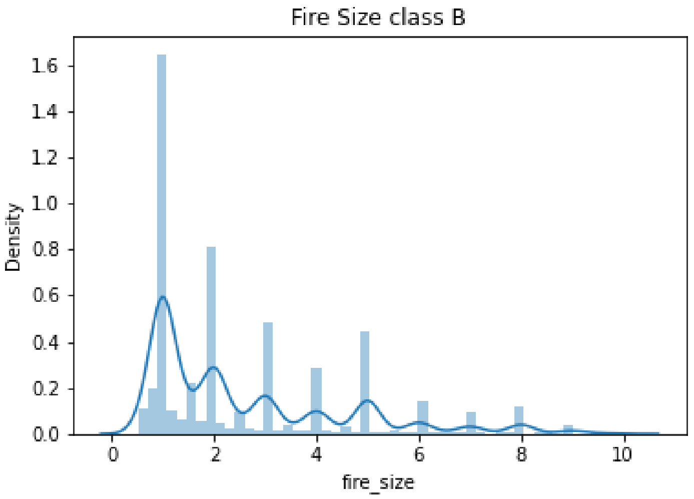
- fire_size — size of fire (in acres)
- Temp_pre_7 — average temperature 7 days before the fire was discovered (in degress Celsius)
- Hum_pre_7 — average humidity 7 days before the fire was discovered (in %)

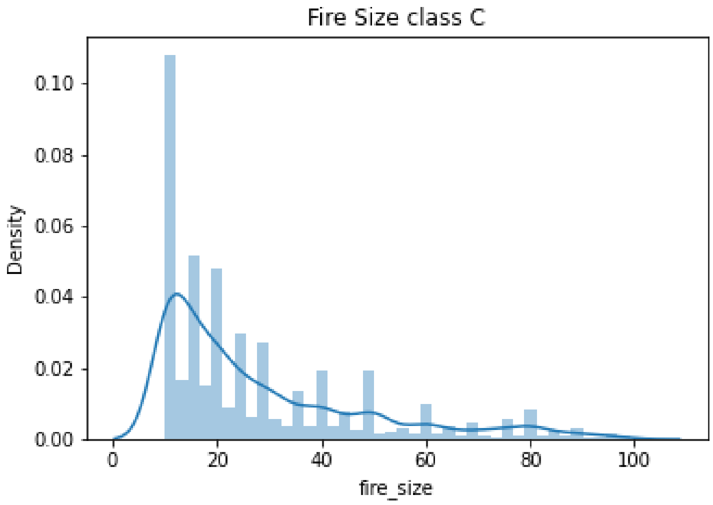
- Wind_pre_7 — average wind speed 7 days before the fire was discovered (in m/s)

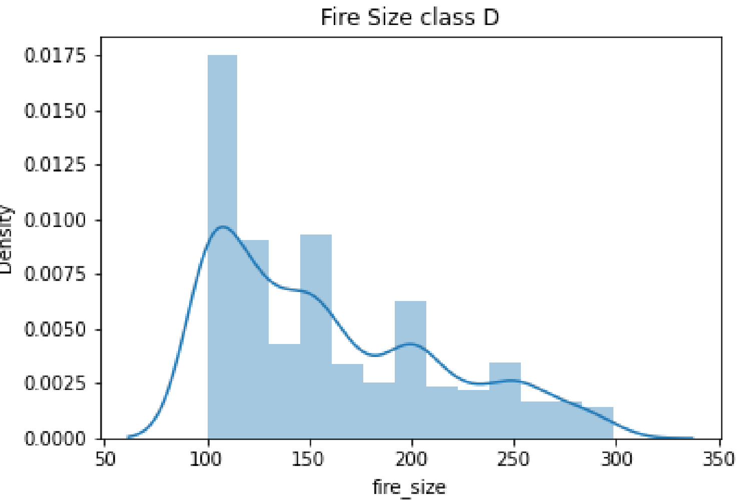
Fire_size we model for each class (how big is the fire per some class), whereas temperature, humidity, wind we model using the whole dataset.

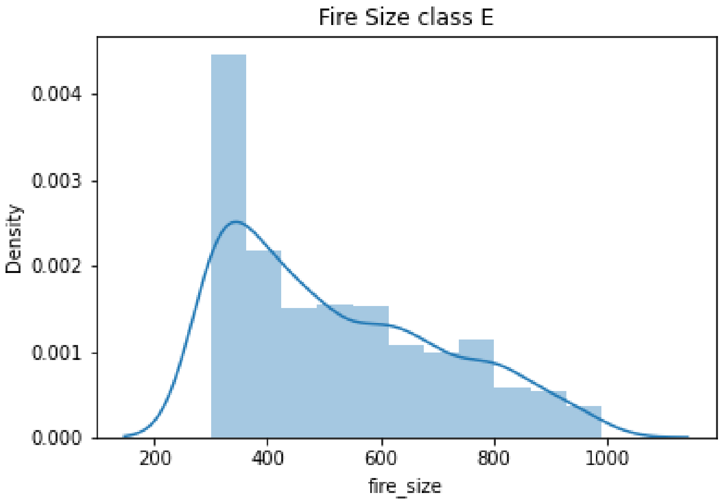
1.1 Modelling fire_size from the perspective of the fire_size class:

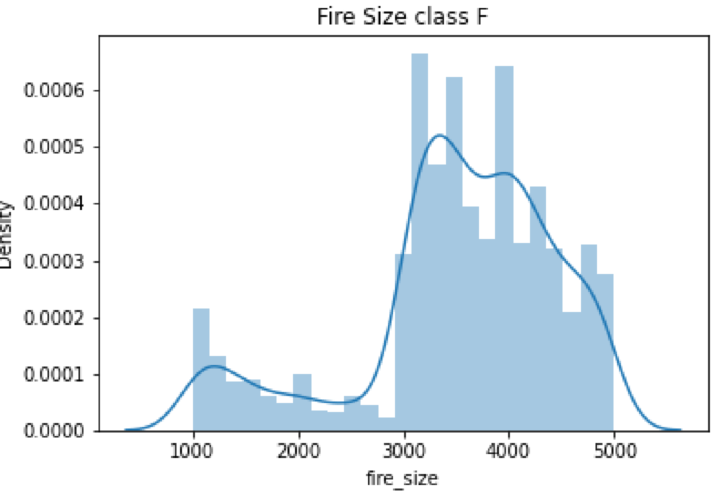
For fire_size we decided to take only one class as an example — class B. This is the class where the most data is located and the probabilities are highest. For other classes we decided to show just how KDE and histograms are distributed..

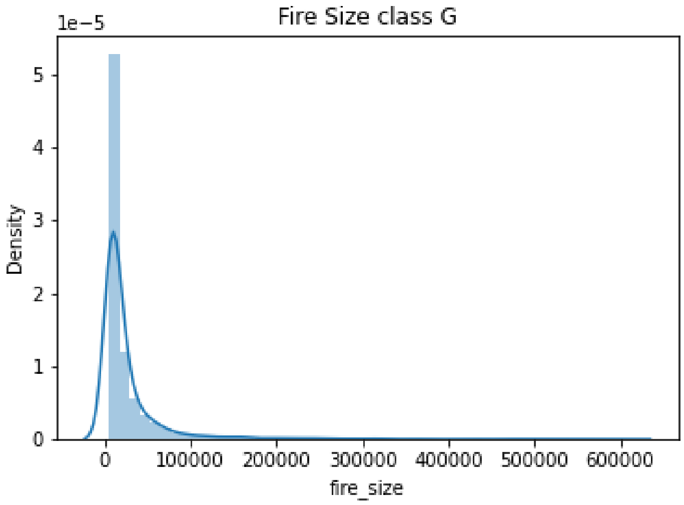


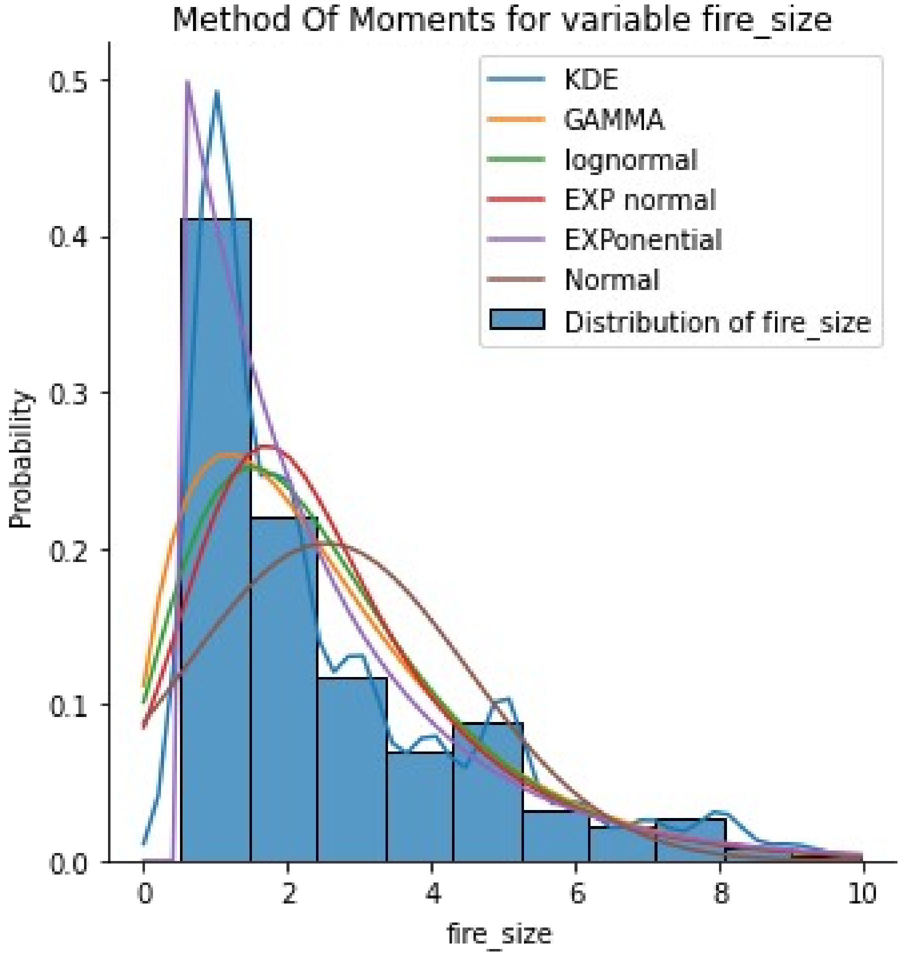












From the diagram we can conclude that the **exponential** distribution seems to resemble the KDE most. The next one looking good by eye is **gamma** or **exponnorm.**

When it comes to the tests:

FOR gamma: Kolmogorov-Smirnoff test result KstestResult(statistic=0.14, pvalue=0.7166468440414822), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.11060000000000159, pvalue=0.5441717496790305)

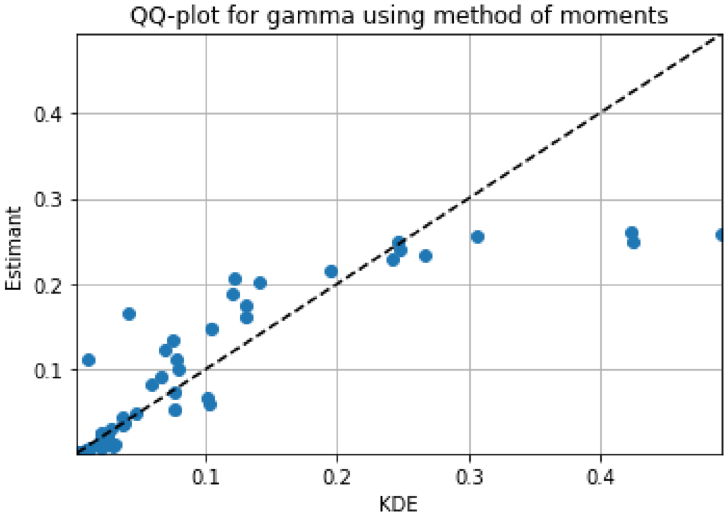
FOR lognorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.14, pvalue=0.7166468440414822), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.1274000000000015, pvalue=0.47358163672801634)

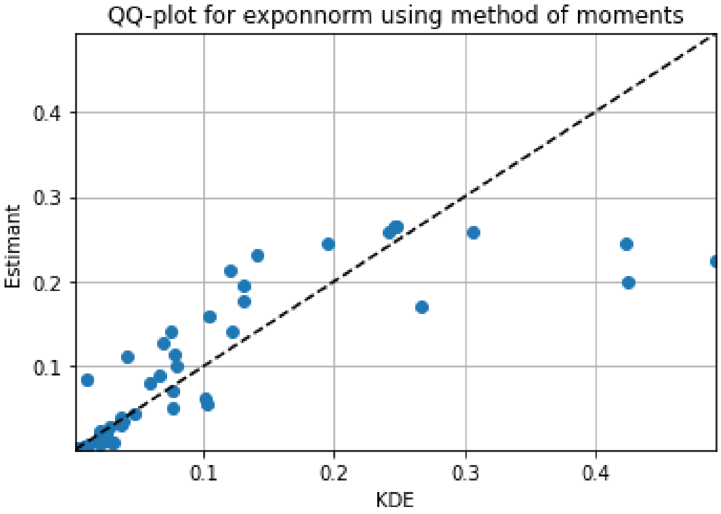
FOR exponnorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.14, pvalue=0.7166468440414822), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.1034000000000006, pvalue=0.578055087825003)

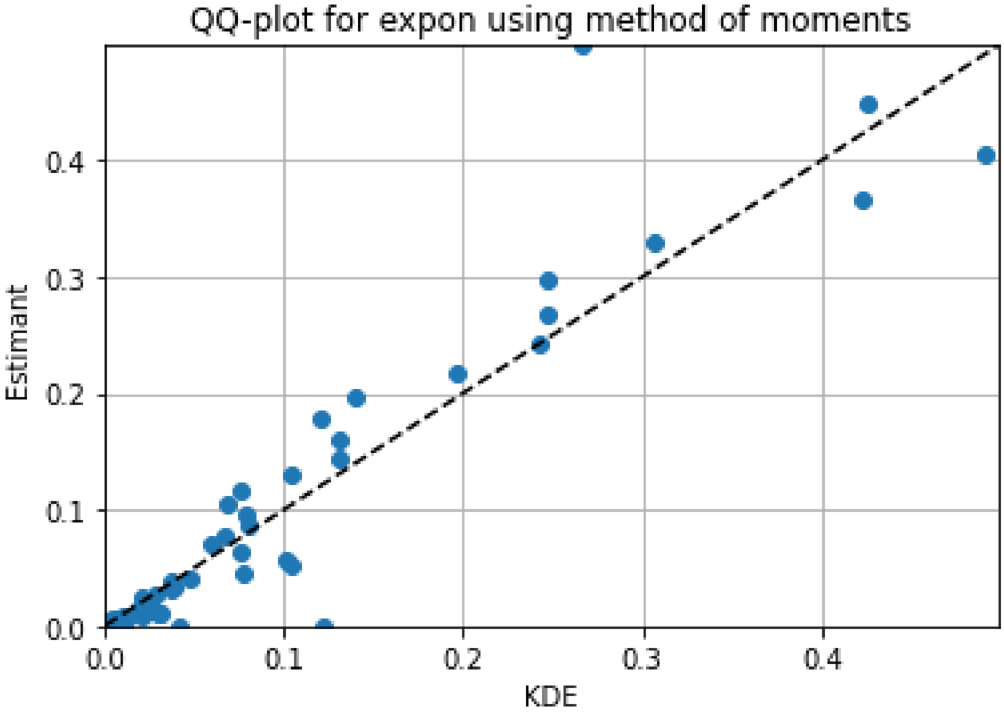
FOR expon: Kolmogorov-Smirnoff test result KstestResult(statistic=0.18, pvalue=0.3959398631708505), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.15020000000000167, pvalue=0.39436389982166364)

FOR norm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.18, pvalue=0.3959398631708505), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.3038000000000025, pvalue=0.13271570041792946)

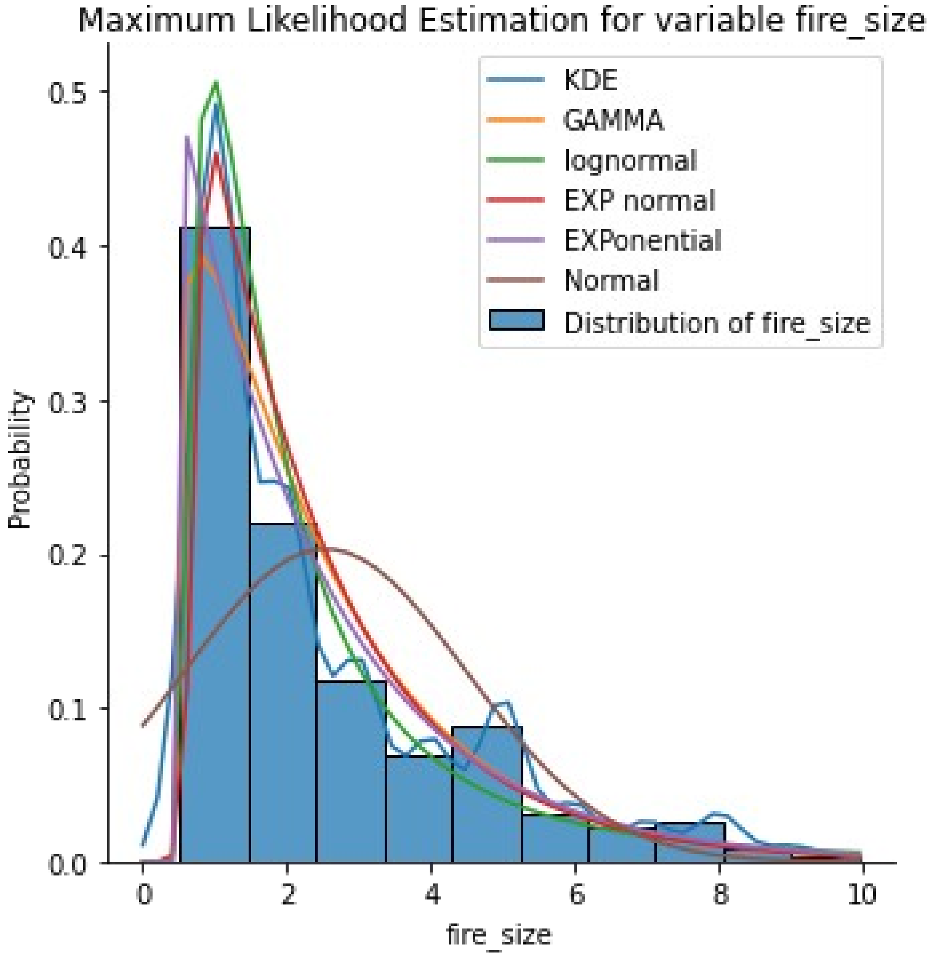
Based on the results we can say that the **gamma** and **exponnorm** are the best estimators.







From Qqplot we can conclude that **expon** is better than the remaining ones. Nevertheless we can spot that all of the distributions seem to fit well the distribution. We can take the **exponential** as the p-value for both tests is bigger than 0.05.



For the Maximum likelihood method (MLE) it is far harder to distinguish the best distribution, as both **exponential, expnormal and lognormal** fit the graph nicely.

Test results:

FOR gamma: Kolmogorov-Smirnoff test result KstestResult(statistic=0.18, pvalue=0.3959398631708505), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.17220000000000013, pvalue=0.3325993319259639)

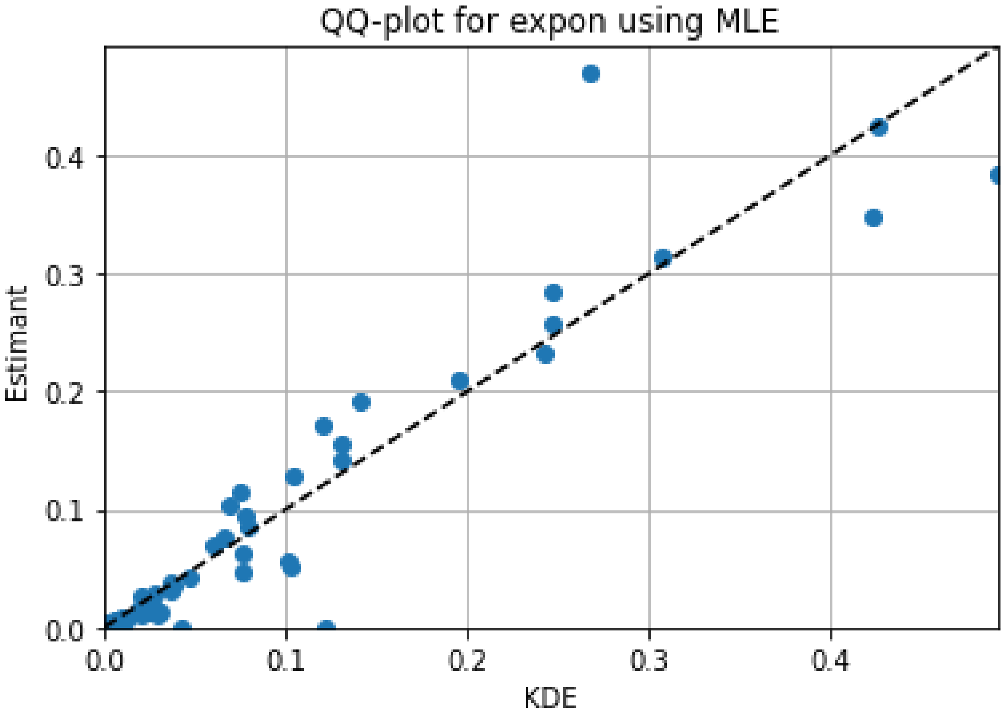
FOR lognorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.22, pvalue=0.17858668181221732), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.2582000000000022, pvalue=0.1796456070337744)

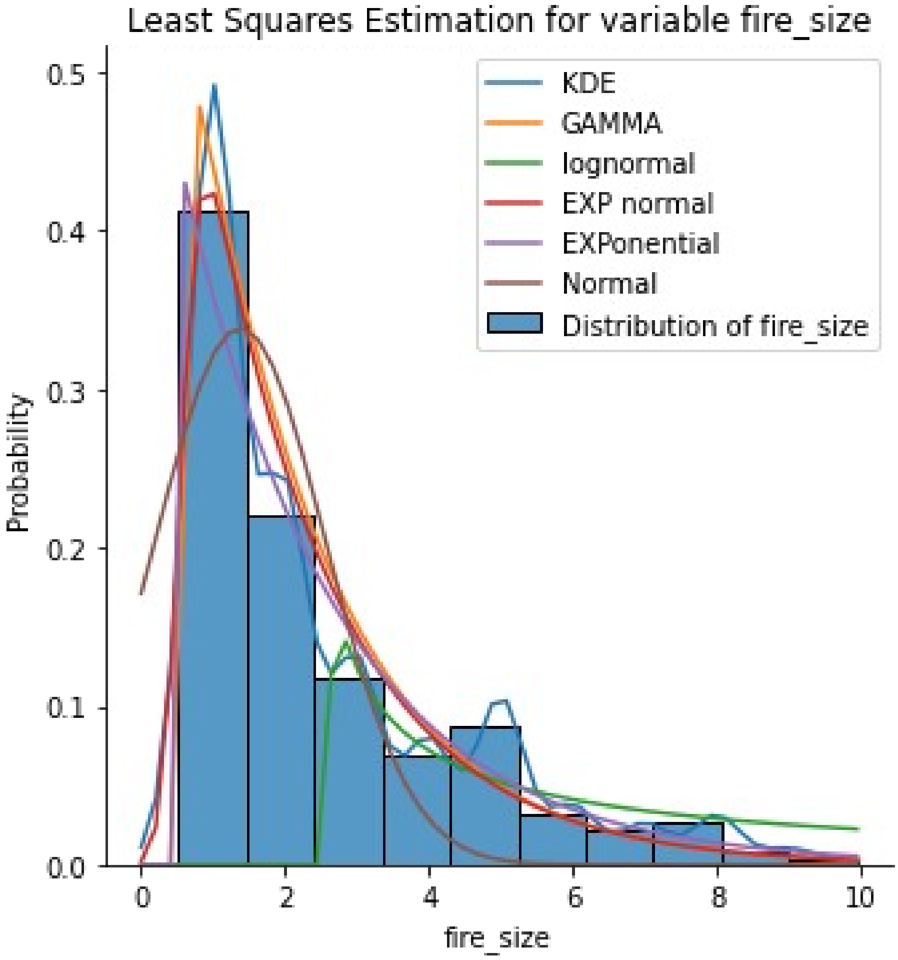
FOR exponnorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.2, pvalue=0.2719135601522248), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.19180000000000064, pvalue=0.2871733055656991)

FOR expon: Kolmogorov-Smirnoff test result KstestResult(statistic=0.16, pvalue=0.5486851446031328), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.1258000000000017, pvalue=0.4798229336666382)

FOR norm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.18, pvalue=0.3959398631708505), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.3038000000000025, pvalue=0.13271570041792946)

Based on tests we can easily conclude that **expon** fits the distribution best.


QQ-plot for expon using MLE
Estimant
KDE
0.0
0.1
0.2
0.3
0.4
0.0
0.1
0.2
0.3
0.4



Similarly as in the previous example, it is hard to assess which is the best, but gamma seems to best fit the diagram.

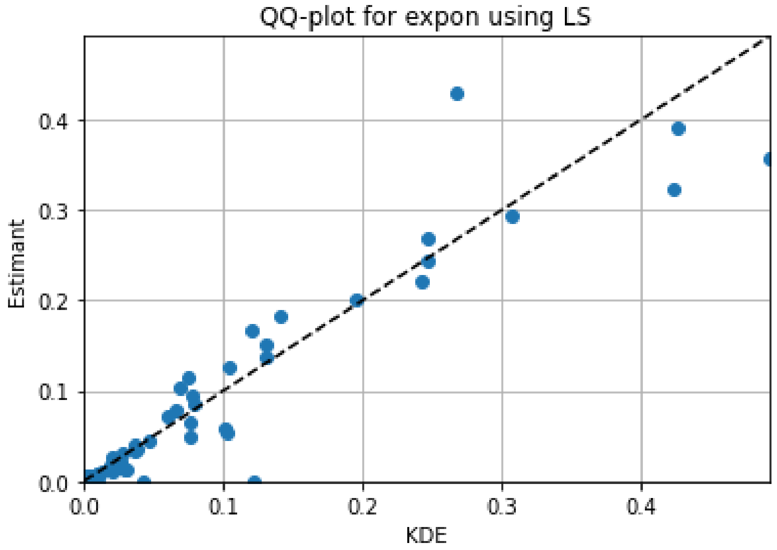
FOR gamma: Kolmogorov-Smirnoff test result KstestResult(statistic=0.22, pvalue=0.17858668181221732), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.2441999999999993, pvalue=0.19770002696435263)

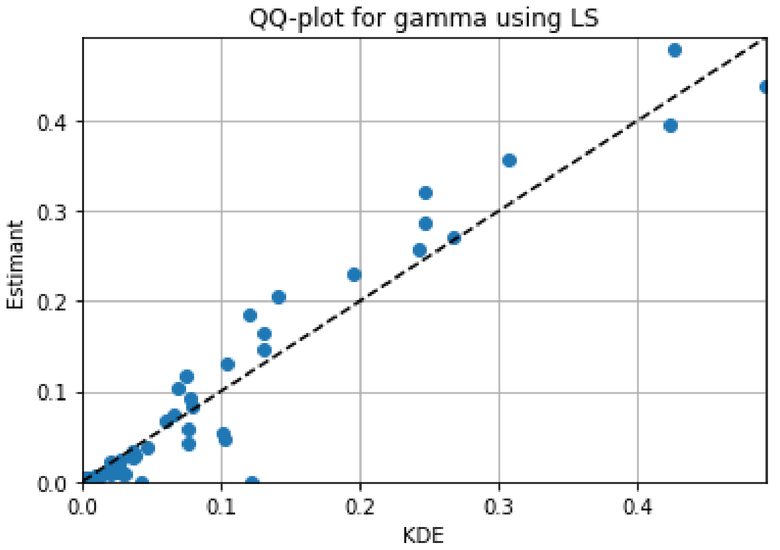
FOR lognorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.32, pvalue=0.011511738725894704), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.8426000000000009, pvalue=0.005732438971866705)

FOR exponnorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.18, pvalue=0.3959398631708505), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.1518000000000015, pvalue=0.3894283242393094)

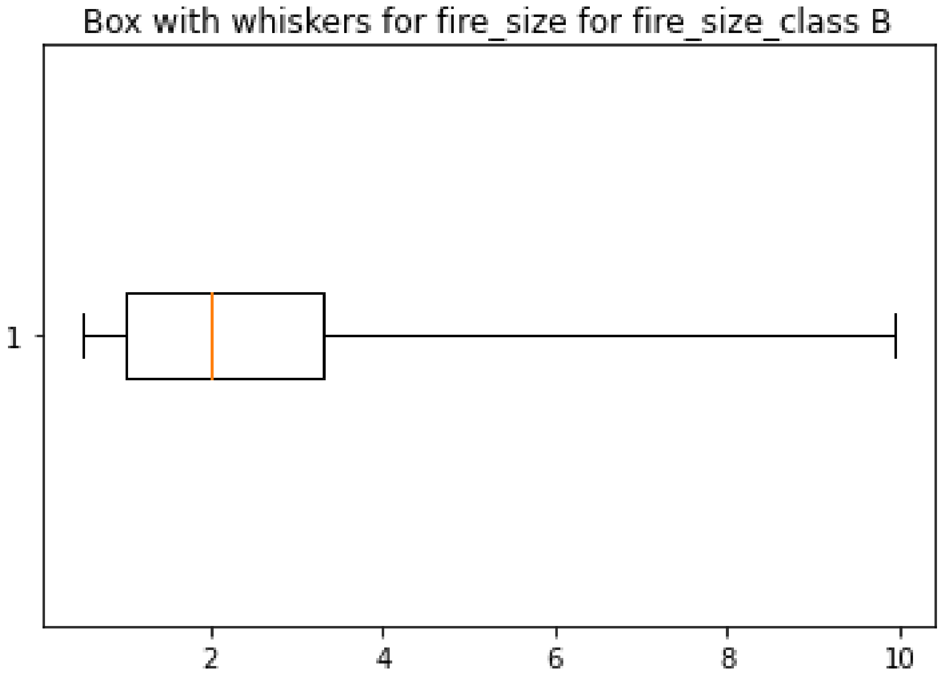
FOR expon: Kolmogorov-Smirnoff test result KstestResult(statistic=0.16, pvalue=0.5486851446031328), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.09500000000000242, pvalue=0.6205176723736756)

FOR norm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.48, pvalue=1.3867885687360081e-05), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=1.5122, pvalue=0.0001575951935046671)

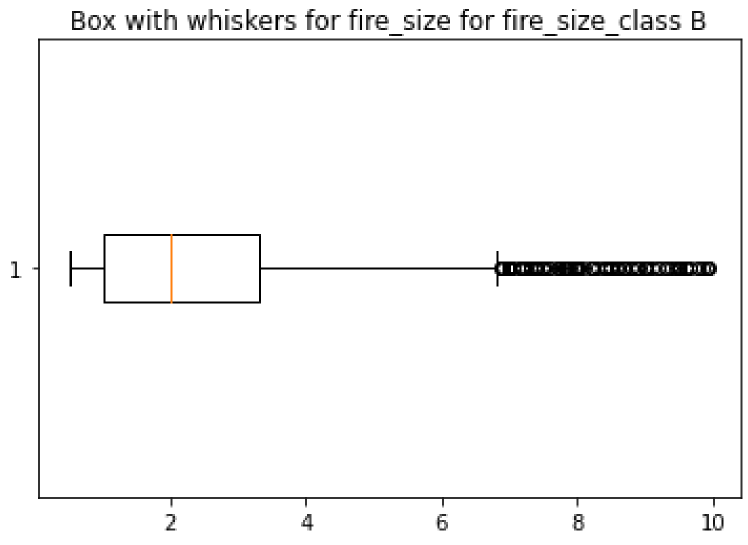




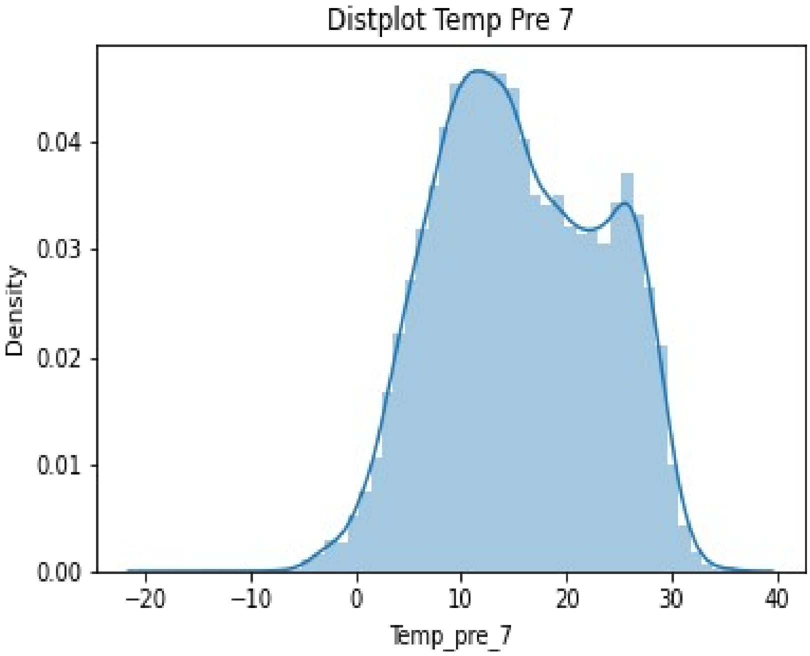
According to the tests, both expon and gamma can be taken into account. On the qq-plot **gamma** plot looks a bit better though.

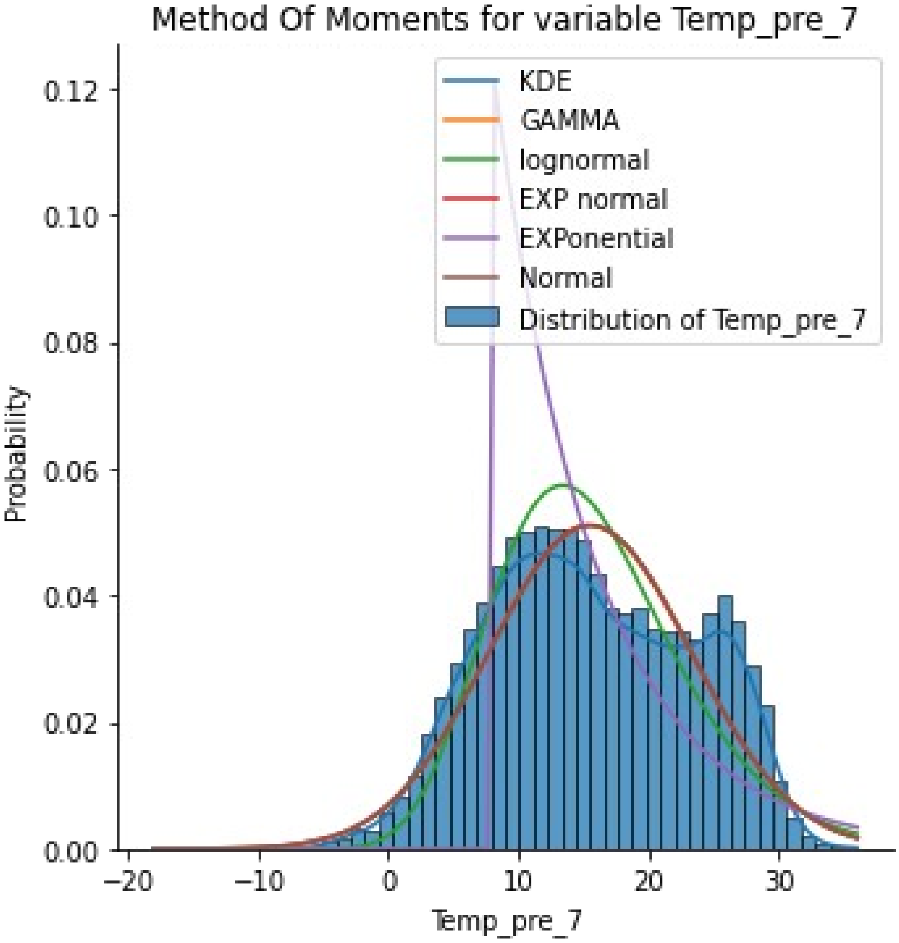


Here we decided to change the fire_size IQR. This is because there were a lot of outliers. Nevertheless, they cannot be treated as outliers as they are within the range of fire_size values for class B. The default IQR would be:



1.2 Modelling Temp_pre_7:





By looking at the diagram one can conclude that either **lognorm** or **exponorm** fit the data best.

FOR gamma: Kolmogorov-Smirnoff test result KstestResult(statistic=0.12, pvalue=0.469506448503778), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.1956999999999951, pvalue=0.27749429769042333)

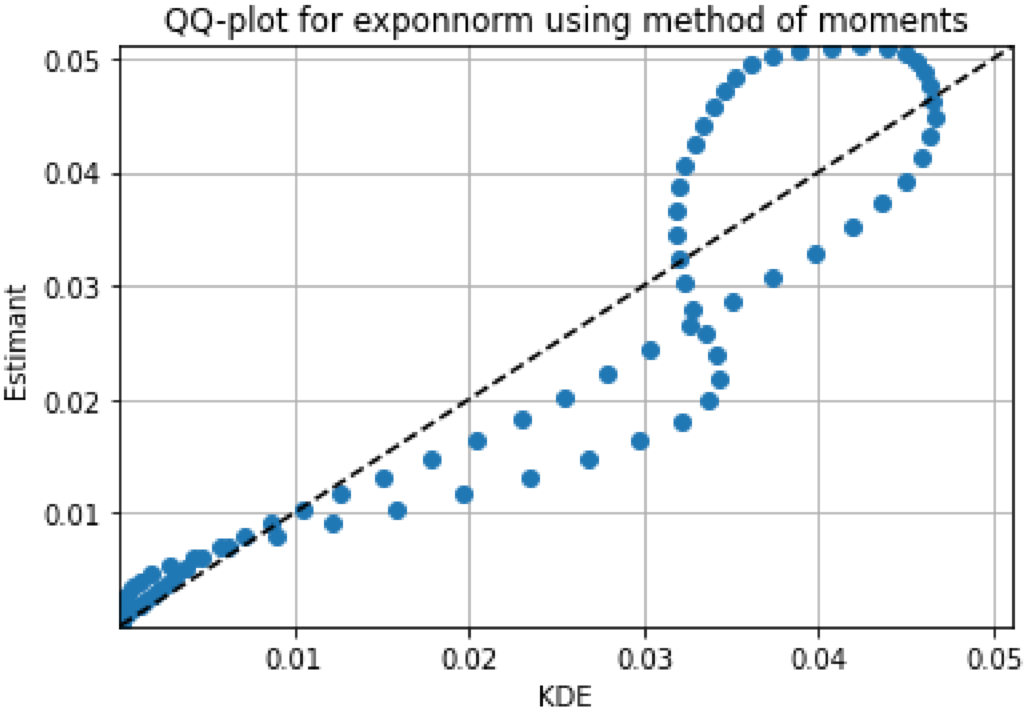
FOR lognorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.23, pvalue=0.009878183186176536), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.42959999999999354, pvalue=0.06048196983166687)

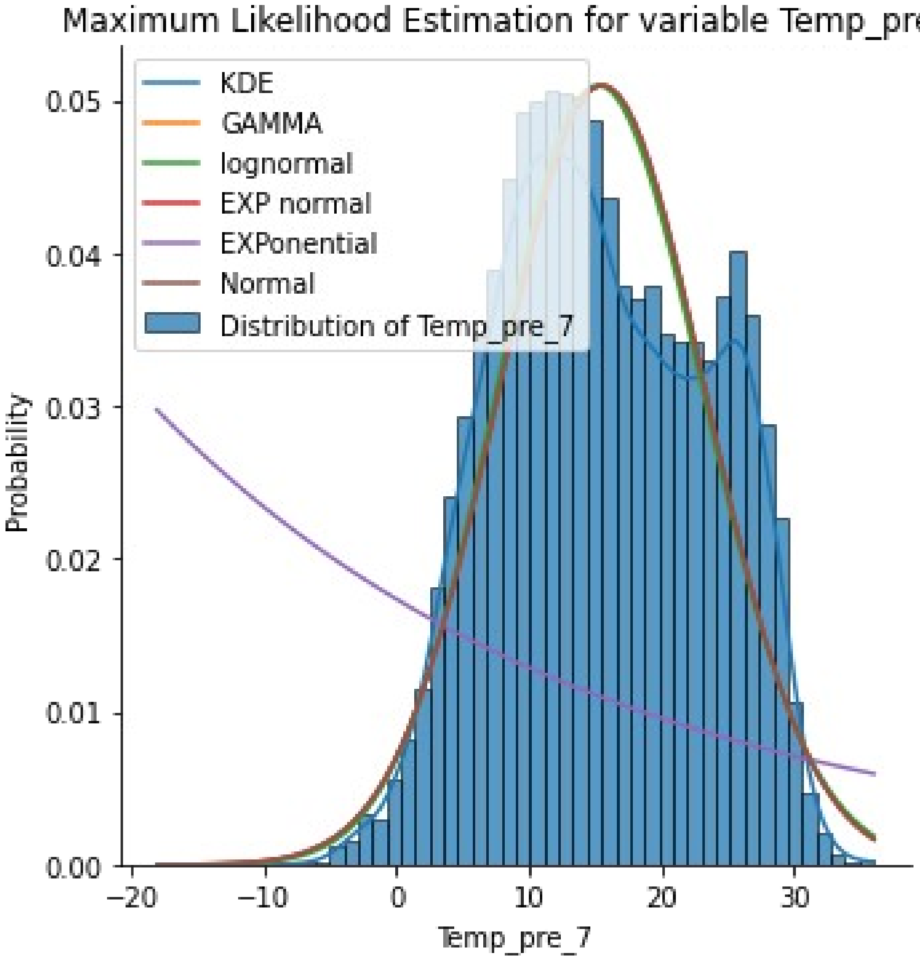
FOR exponnorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.12, pvalue=0.469506448503778), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.19999999999999574, pvalue=0.2689298575502499)

FOR expon: Kolmogorov-Smirnoff test result KstestResult(statistic=0.48, pvalue=8.448372017533173e-11), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=2.7958, pvalue=2.08176859972653e-07)

FOR norm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.12, pvalue=0.469506448503778), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.21640000000000015, pvalue=0.23902942010749217)

By summing up the results **exponnorm** seems to be the best here.





Here both **exponormal** and **lognormal** seem to fit the diagram.

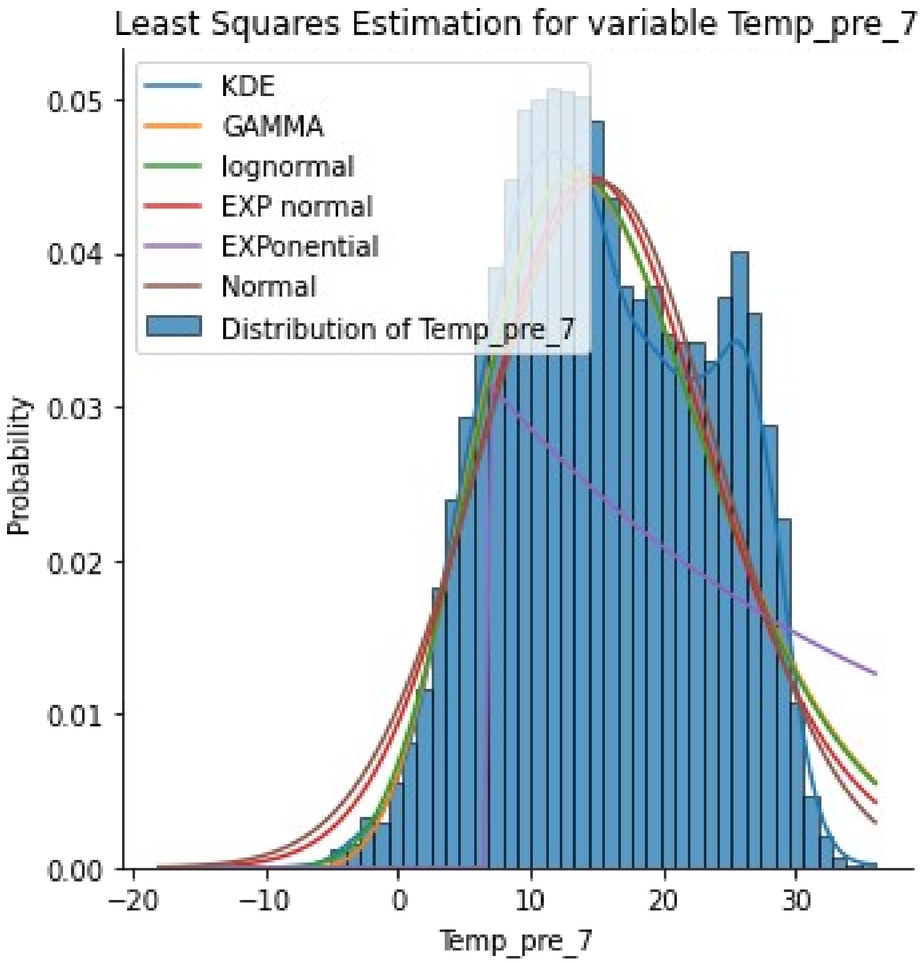
FOR gamma: Kolmogorov-Smirnoff test result KstestResult(statistic=0.12, pvalue=0.469506448503778), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.19589999999999463, pvalue=0.2770887835182927)

FOR lognorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.12, pvalue=0.469506448503778), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.1862999999999957, pvalue=0.297381944024553)

FOR exponnorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.12, pvalue=0.469506448503778), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.21640000000000015, pvalue=0.23902942010749217)

FOR expon: Kolmogorov-Smirnoff test result KstestResult(statistic=0.43, pvalue=1.1151678185620634e-08), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=2.904299999999999, pvalue=1.1949203970740285e-07)

FOR norm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.12, pvalue=0.469506448503778), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.21640000000000015, pvalue=0.23902942010749217)



From the diagram we can conclude that **lognormal, exponormal** and **normal** distributions fit the data.

FOR gamma: Kolmogorov-Smirnoff test result KstestResult(statistic=0.2, pvalue=0.03638428787491733), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.28429999999999467, pvalue=0.1502961971174701)

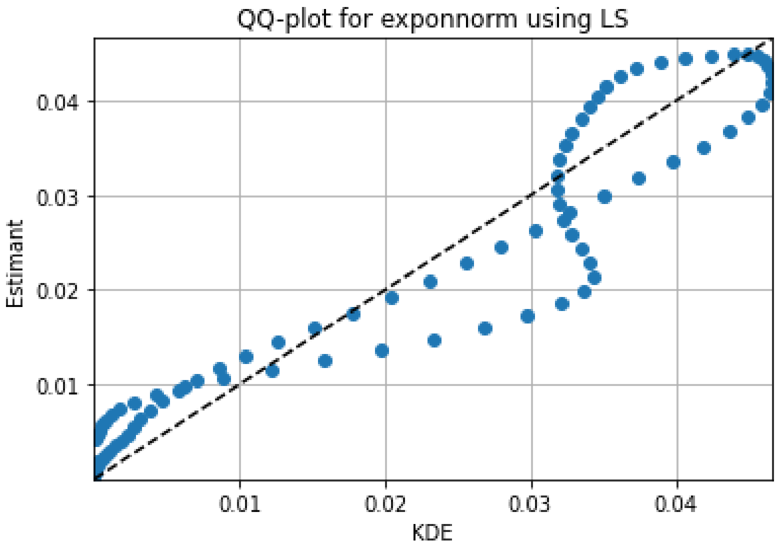
FOR lognorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.15, pvalue=0.21117008625127576), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.1903999999999968, pvalue=0.28850410012600614)

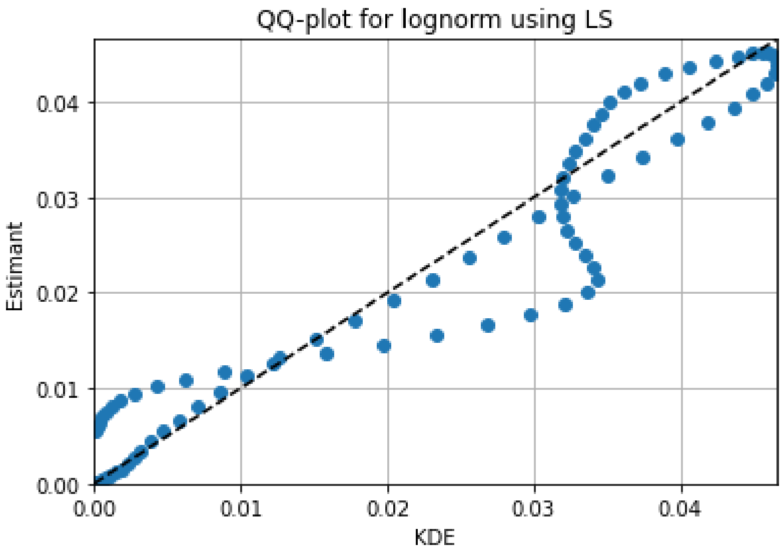
FOR exponnorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.14, pvalue=0.2819416298082479), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.3188999999999993, pvalue=0.12002998141084942)

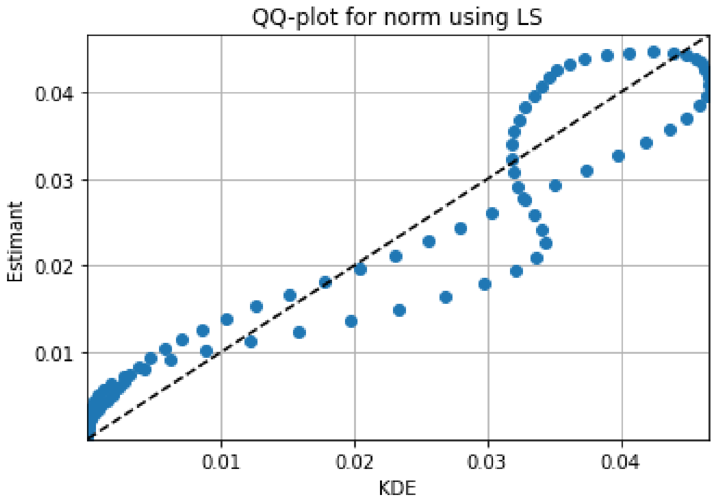
FOR expon: Kolmogorov-Smirnoff test result KstestResult(statistic=0.46, pvalue=6.422179651064002e-10), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=3.3373749999999944, pvalue=1.3116873032181786e-08)

FOR norm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.18, pvalue=0.07822115797841851), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.4228999999999985, pvalue=0.0629728394780188)

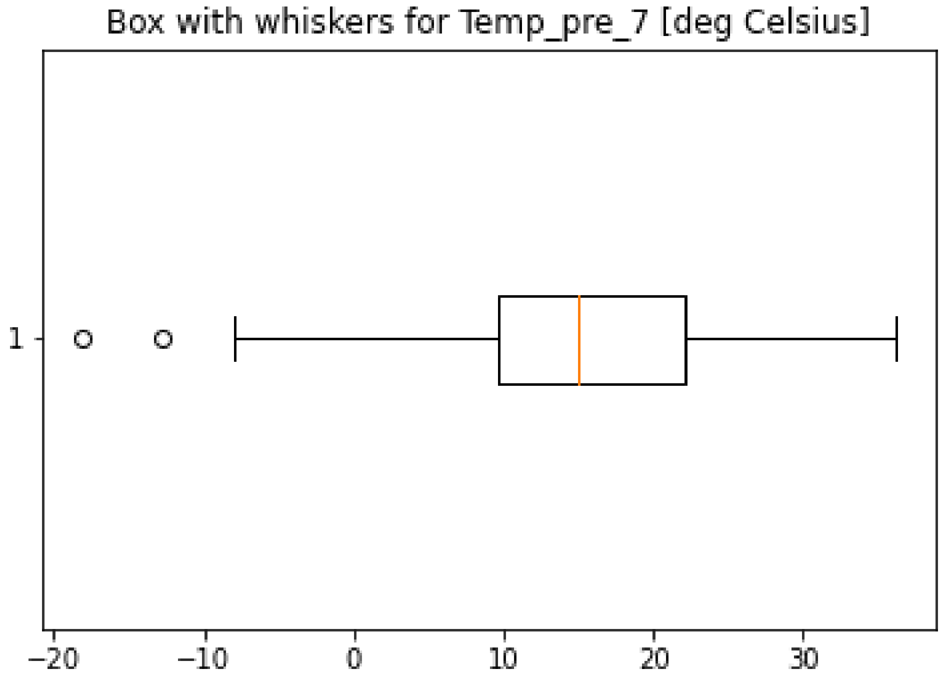
Both three distributions are not rejected by the tests. For them the qq-plots are:



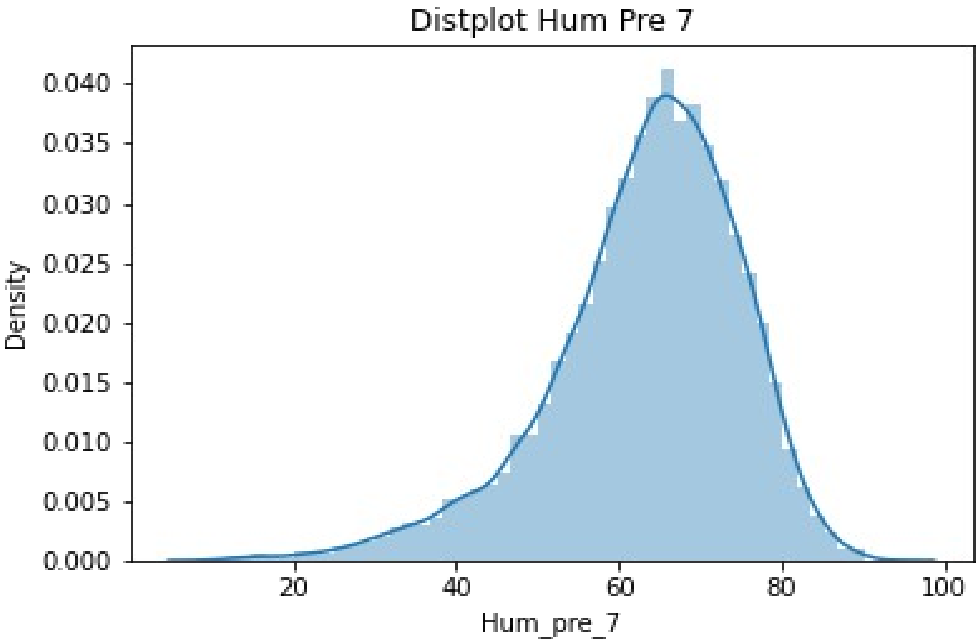


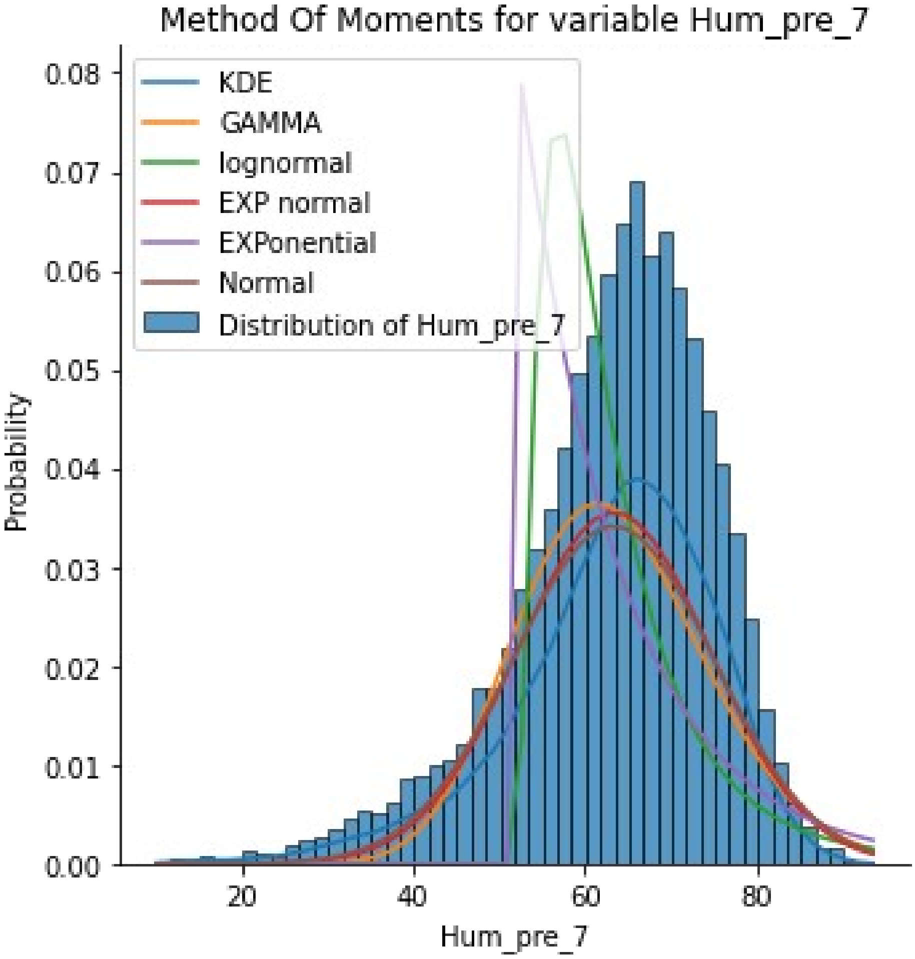


As we can see all of those distributions do match the real data. Loking at the graph however and qq-plots the **lognorm** is probably the best depiction of the processes.



1.3 Modelling Hum_pre_7





Based on the diagram we can see that **gamma** and **expnormal** seem to best fit the distribution. The results of tests are:

FOR gamma: Kolmogorov-Smirnoff test result KstestResult(statistic=0.24, pvalue=0.11238524845512393), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.20139999999999958, pvalue=0.2676578770784733)

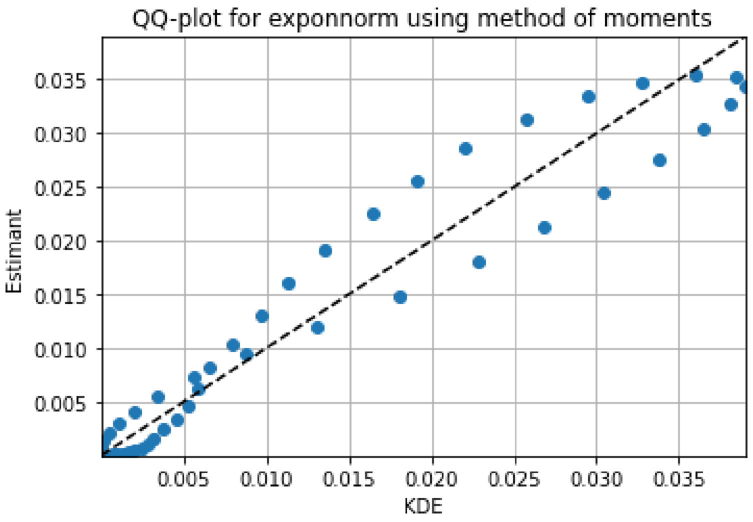
FOR lognorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.5, pvalue=4.8075337049514946e-06), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=1.5153999999999996, pvalue=0.0001549589235869675)

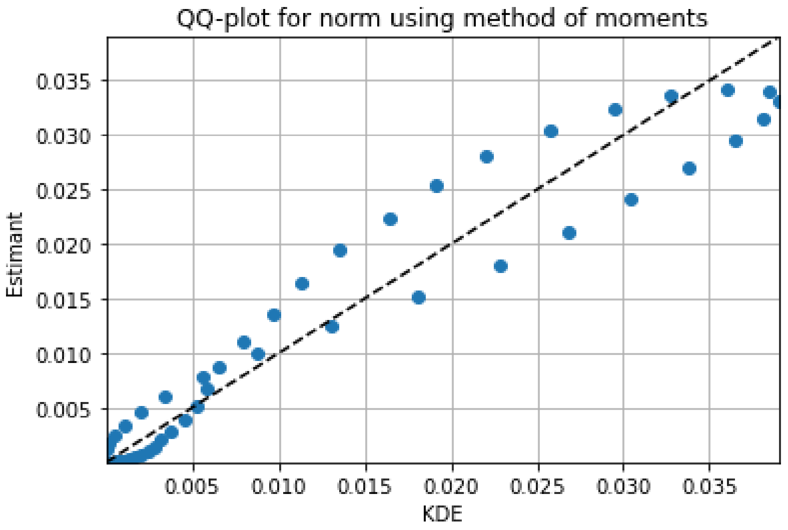
FOR exponnorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.16, pvalue=0.5486851446031328), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.09660000000000224, pvalue=0.6121826683772482)

FOR expon: Kolmogorov-Smirnoff test result KstestResult(statistic=0.5, pvalue=4.8075337049514946e-06), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=1.413800000000002, pvalue=0.00026501909924037115)

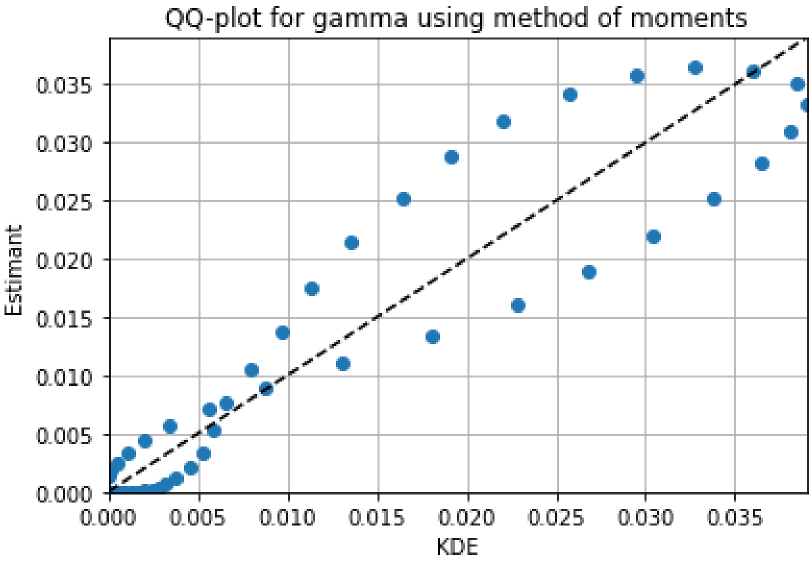
FOR norm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.16, pvalue=0.5486851446031328), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.08539999999999992, pvalue=0.672953892101148)

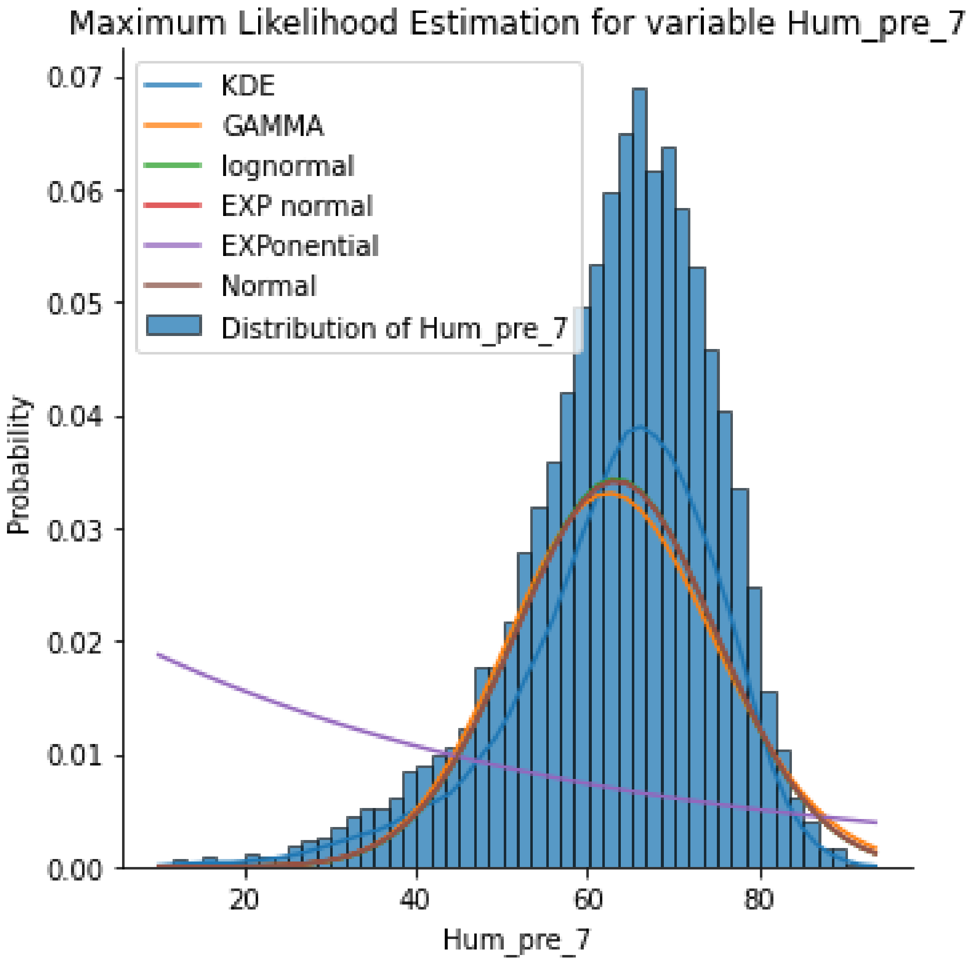
The QQ-plots for the following distributions:





As we can see, based on the results, **exponnorm** will be the best model.





On this diagram we can see that **lognormal, expnormal and normal** are best best here.

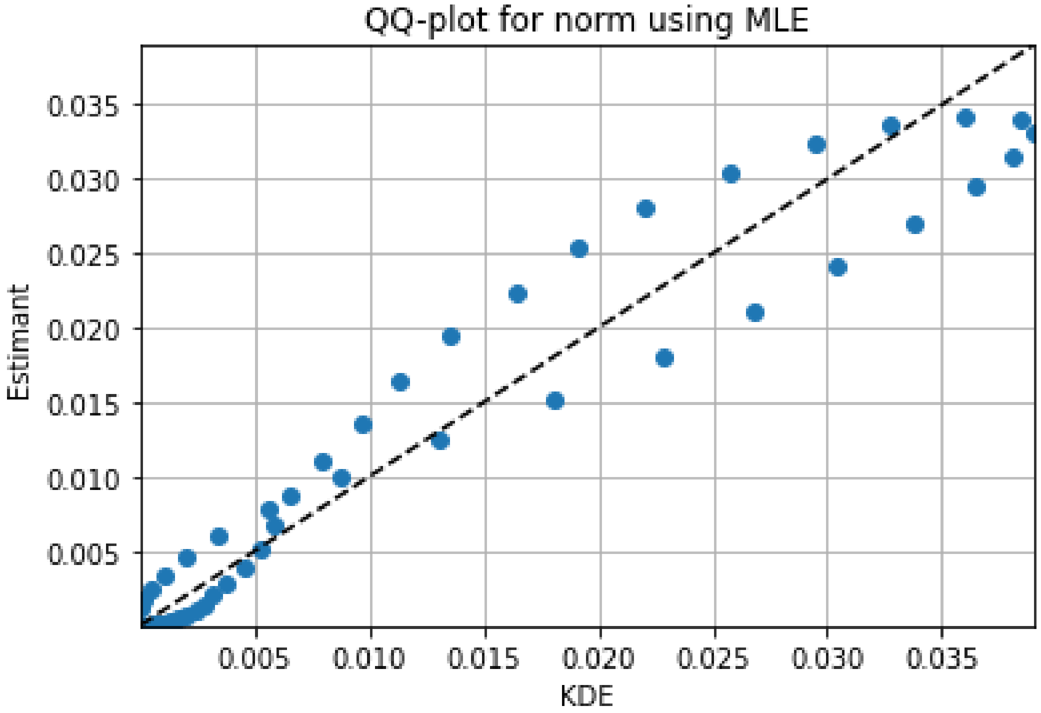
FOR gamma: Kolmogorov-Smirnoff test result KstestResult(statistic=0.16, pvalue=0.5486851446031328), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.10699999999999932, pvalue=0.5608283253009358)

FOR lognorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.16, pvalue=0.5486851446031328), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.08859999999999957, pvalue=0.6550180445705647)

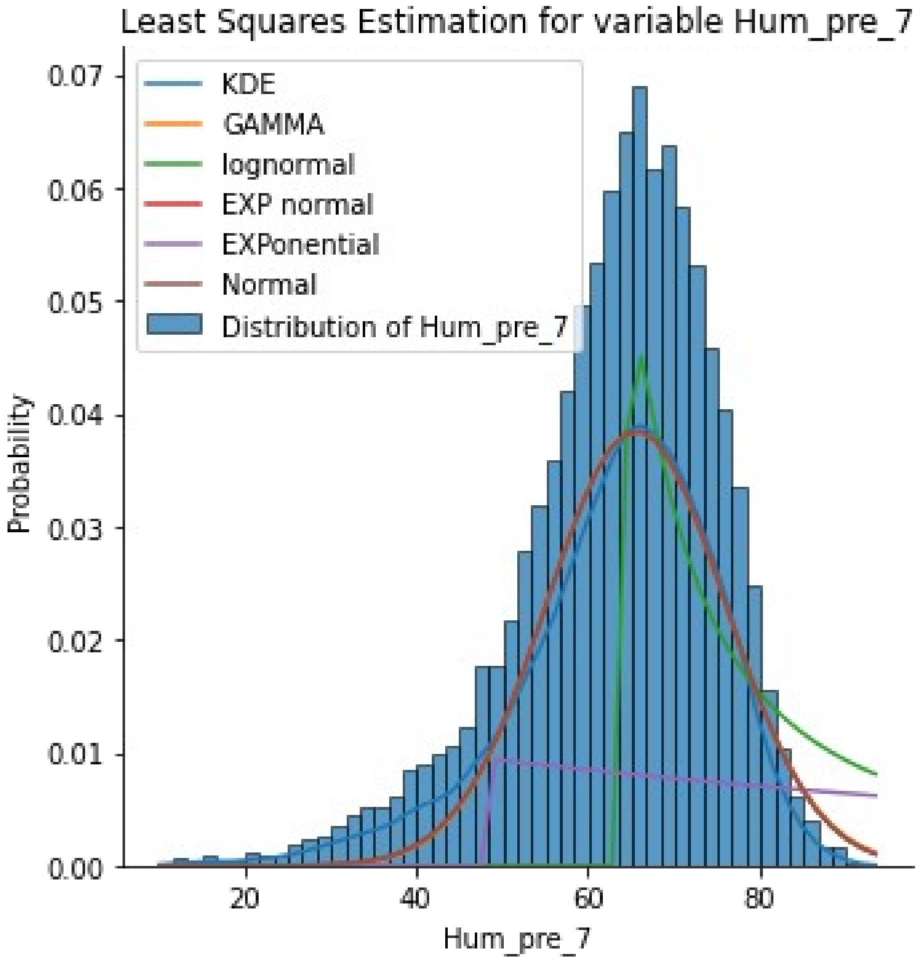
FOR exponnorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.16, pvalue=0.5486851446031328), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.08539999999999992, pvalue=0.672953892101148)

FOR expon: Kolmogorov-Smirnoff test result KstestResult(statistic=0.46, pvalue=3.800827929128319e-05), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=1.2698, pvalue=0.0005693418560481778)

FOR norm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.16, pvalue=0.5486851446031328), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.08539999999999992, pvalue=0.672953892101148)



Based on the general size and look of the distribution one can conclude that it is a **normal** distribution as it is bothnd resembles the data itself.



Here we can notice that KDE seems very similar to **expnormal** and **gamma** and **normal.**

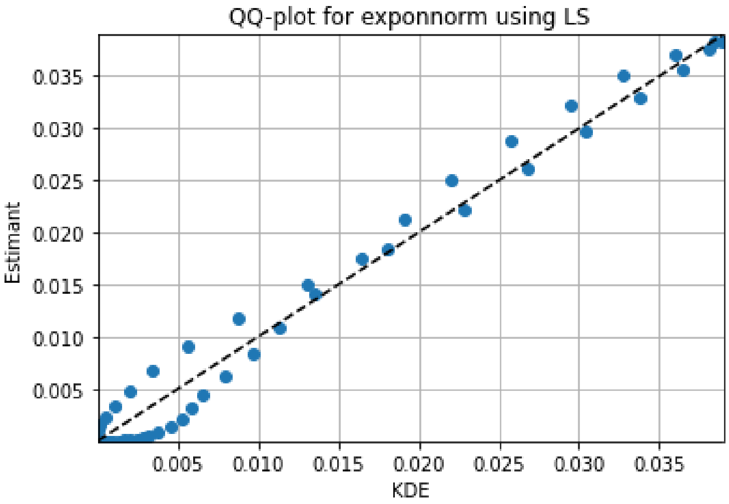
FOR gamma: Kolmogorov-Smirnoff test result KstestResult(statistic=0.24, pvalue=0.11238524845512393), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.2137999999999991, pvalue=0.24474145471620323)

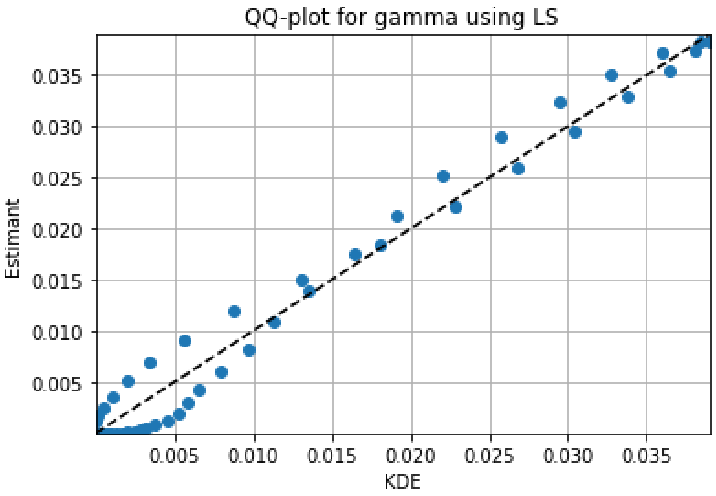
FOR lognorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.64, pvalue=6.078719823015066e-10), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=2.8198000000000008, pvalue=1.7849144673398598e-07)

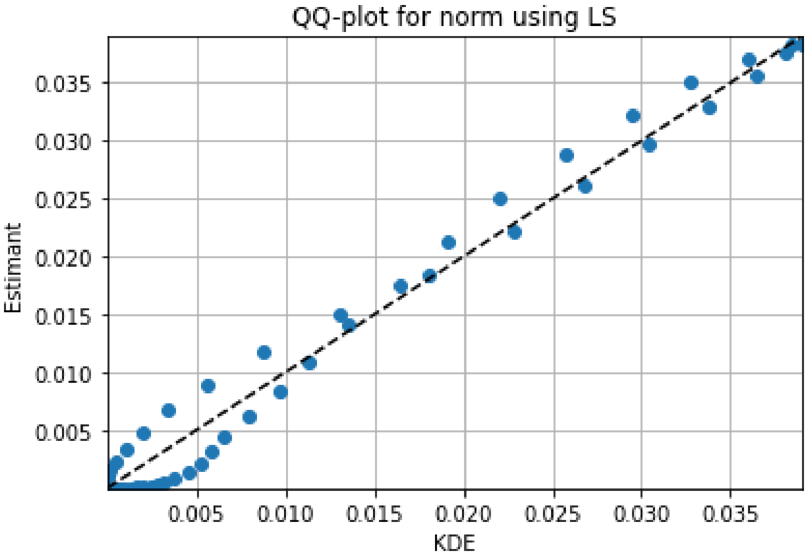
FOR exponnorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.22, pvalue=0.17858668181221732), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.1974000000000018, pvalue=0.27559013347301053)

FOR expon: Kolmogorov-Smirnoff test result KstestResult(statistic=0.46, pvalue=3.800827929128319e-05), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=1.6274000000000015, pvalue=8.596491207235601e-05)

FOR norm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.22, pvalue=0.17858668181221732), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.1974000000000018, pvalue=0.27559013347301053)



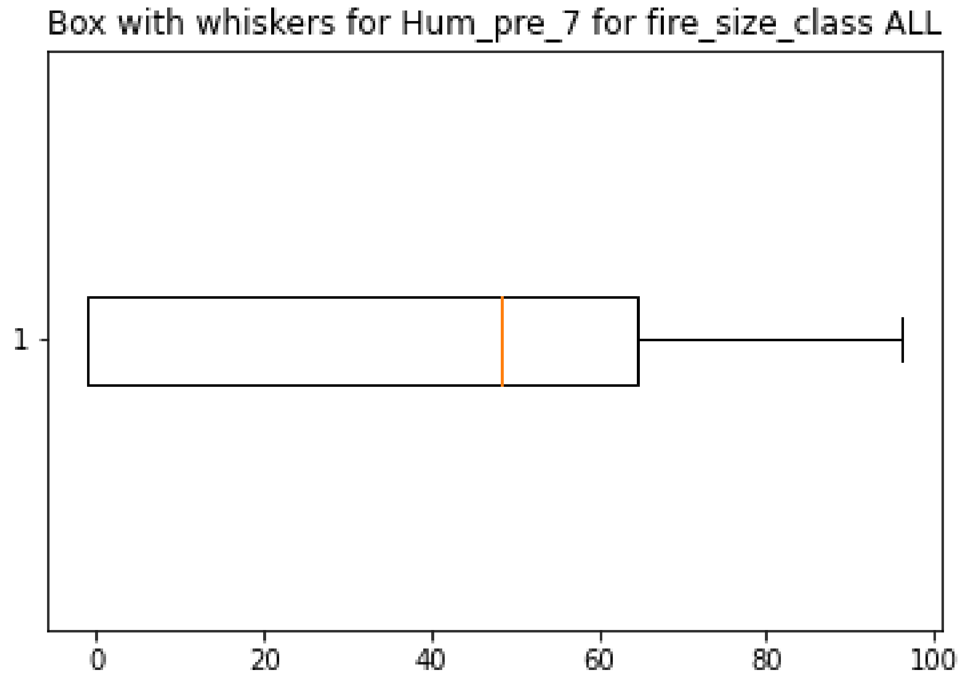




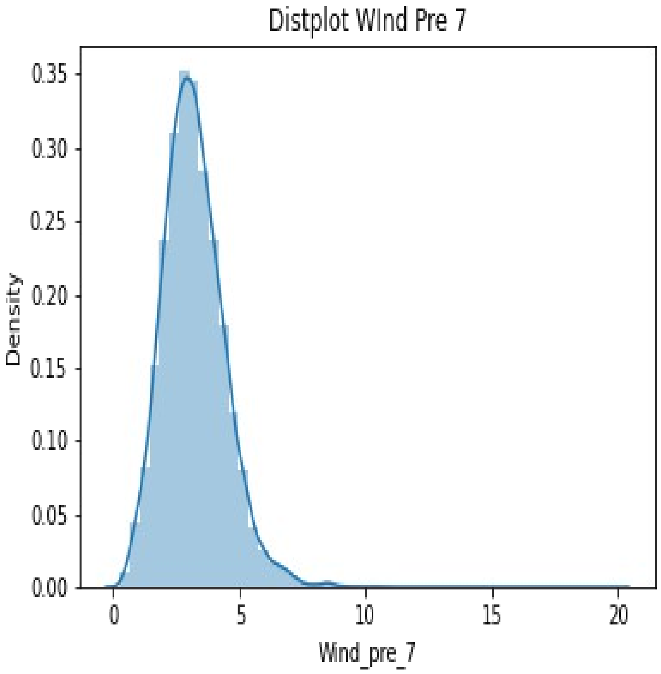
My observations do match the results of the tests as only those 3 distributions can be accepted.

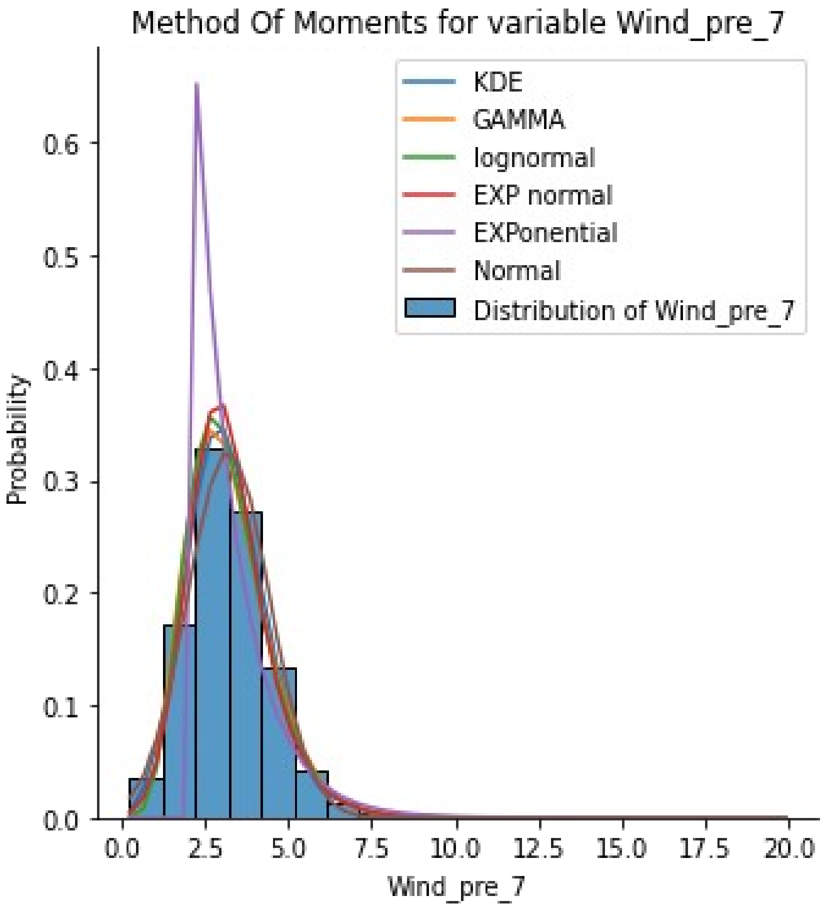
As we can see all three distributions do match the data very well. In my opinion all 3 can be accepted, nevertheless for me either **gamma** or **normal** are the most natural.

Box with whiskers for this variable is:



1.4 Wind_pre_7





As we can see there are many distributions that seem to fit the data at the first glance. Those are **lognormal, expnormal, normal.** Nevertheless, the data seem to be normally distributed.

FOR gamma: Kolmogorov-Smirnoff test result KstestResult(statistic=0.36, pvalue=0.0028349805813203420), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.4578000000000024, pvalue=0.05112092905699761)

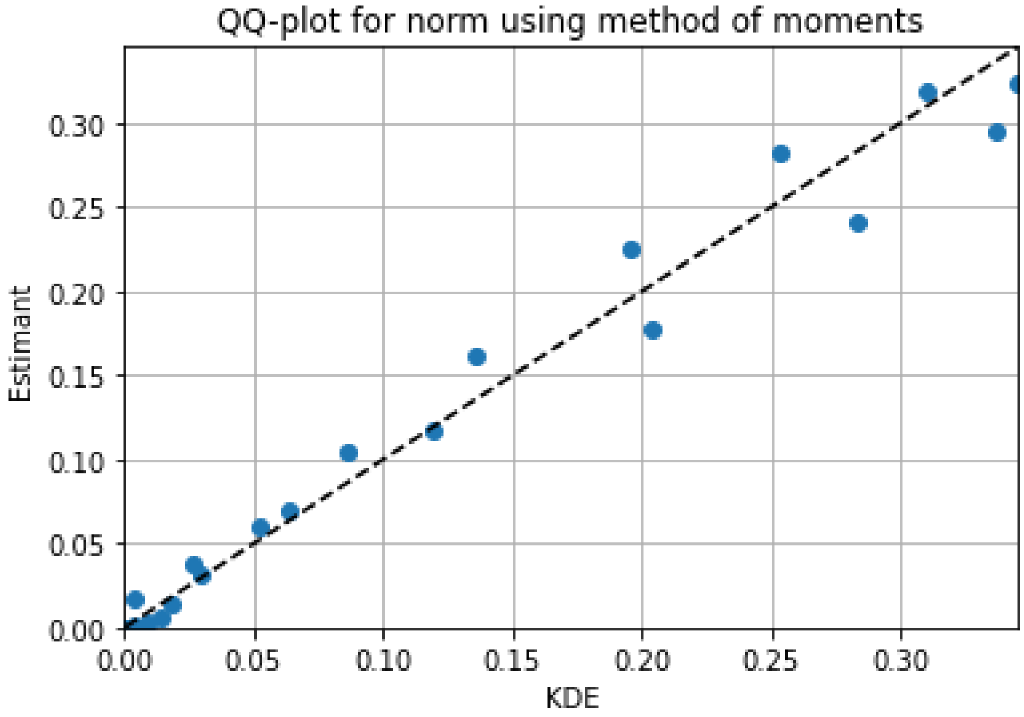
FOR lognorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.38, pvalue=0.0013147736033165794), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.5137999999999998, pvalue=0.036729302938237396)

FOR exponnorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.38, pvalue=0.0013147736033165794), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.5457999999999998, pvalue=0.03048485882684937)

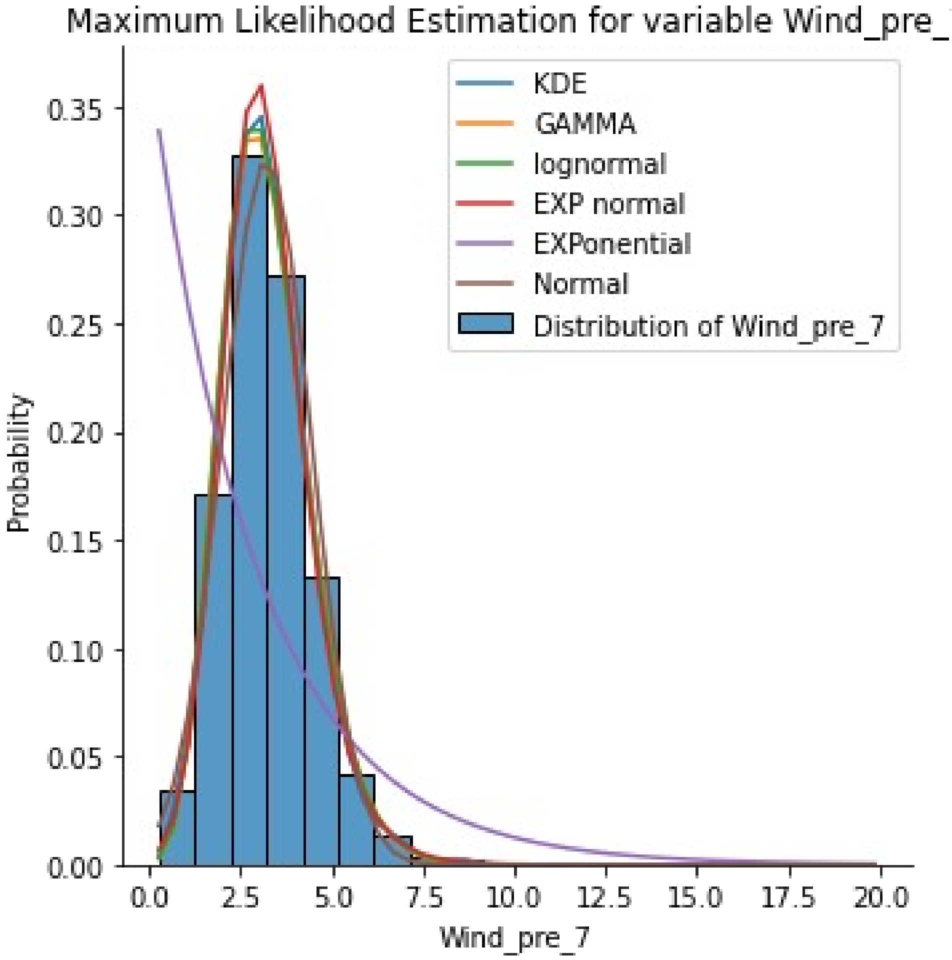
FOR expon: Kolmogorov-Smirnoff test result KstestResult(statistic=0.3, pvalue=0.02170784069014051), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.30020000000000024, pvalue=0.13586505813887295)

FOR norm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.24, pvalue=0.11238524845512393), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.21620000000000061, pvalue=0.2405796888066748)

Amongst the tests only **norm** seem to fit the data.



According to QQ plot the **normal** distribution seems to fit really good.



Here, again, many distributions seem to fit well the data. Nevertheless the distribution seems to be **normal.**

FOR gamma: Kolmogorov-Smirnoff test result KstestResult(statistic=0.34, pvalue=0.005841778142694731), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.41100000000000136, pvalue=0.06775609572335606)

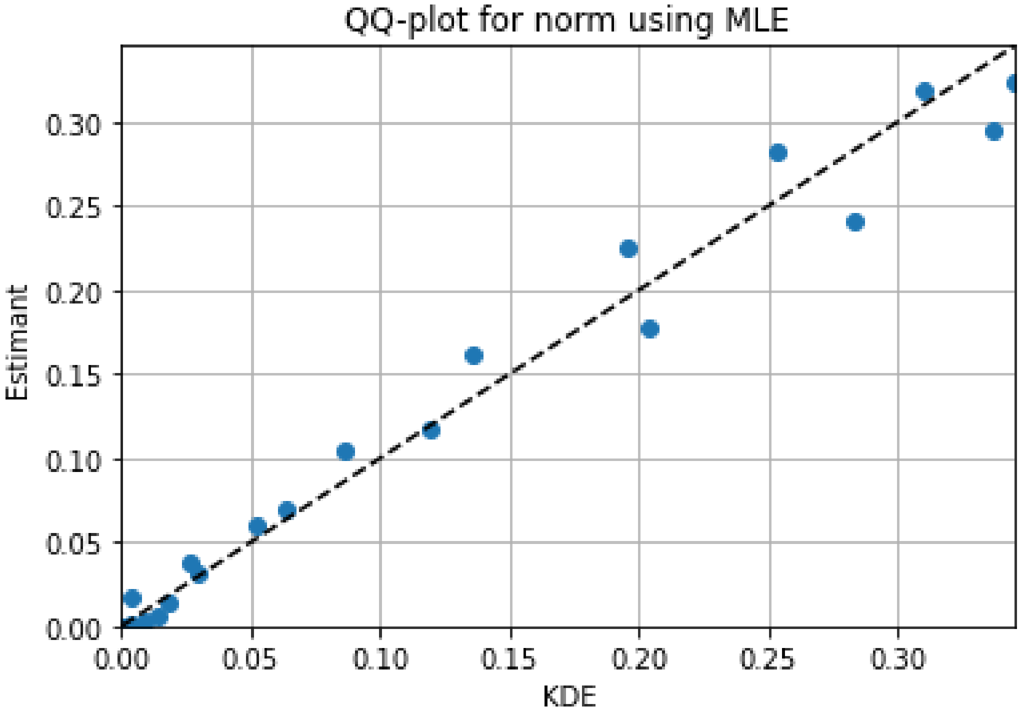
FOR lognorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.36, pvalue=0.0028349805813203342), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.4529999999999994, pvalue=0.052606475413884635)

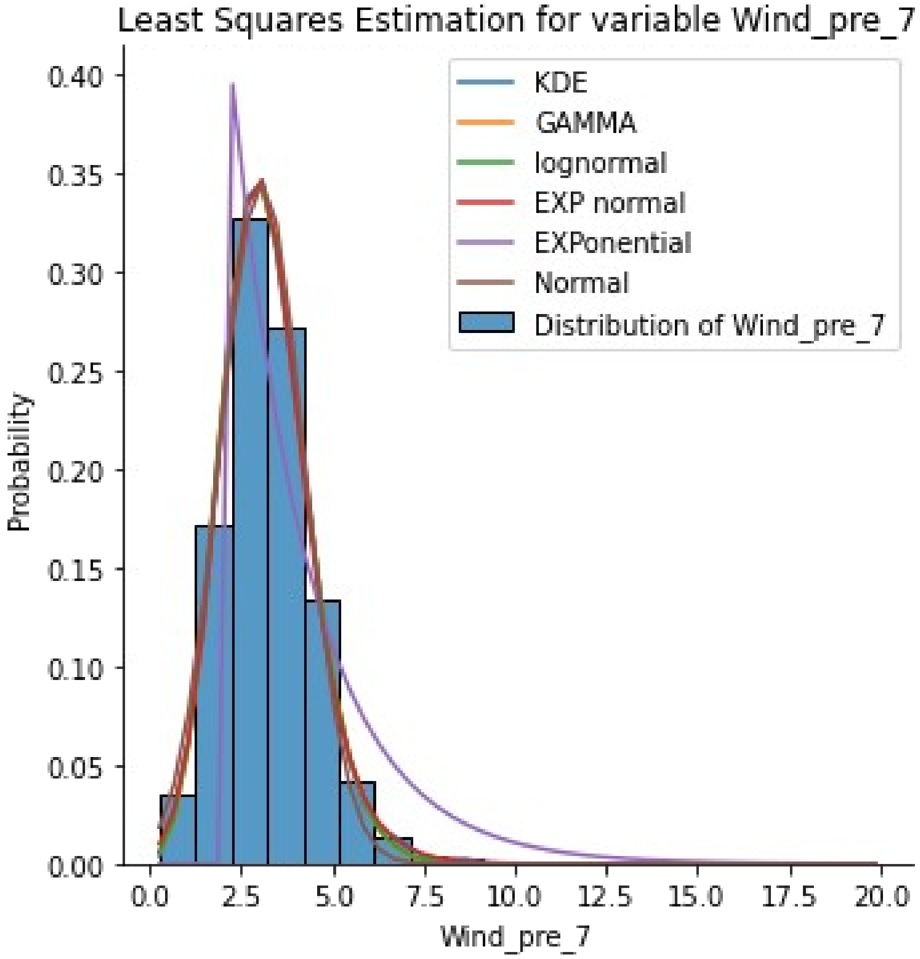
FOR exponnorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.38, pvalue=0.0013147736033165794), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.5226000000000006, pvalue=0.03488859460465488)

FOR expon: Kolmogorov-Smirnoff test result KstestResult(statistic=0.54, pvalue=4.929118631187453e-07), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=1.7570000000000014, pvalue=4.3588816605488745e-05)

FOR norm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.24, pvalue=0.11238524845512393), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.2162000000000061, pvalue=0.2405796888066748)

According to the tests only **normal** distribution fits the data.


QQ-plot for norm using MLE
Estimant
KDE
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.00
0.05
0.10
0.15
0.20
0.25
0.30



Similarly as in the previous examples, **normal distribution should match the best.**

FOR gamma: Kolmogorov-Smirnoff test result KstestResult(statistic=0.32, pvalue=0.011511738725894704), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.3666000000000018, pvalue=0.08903356380455107)

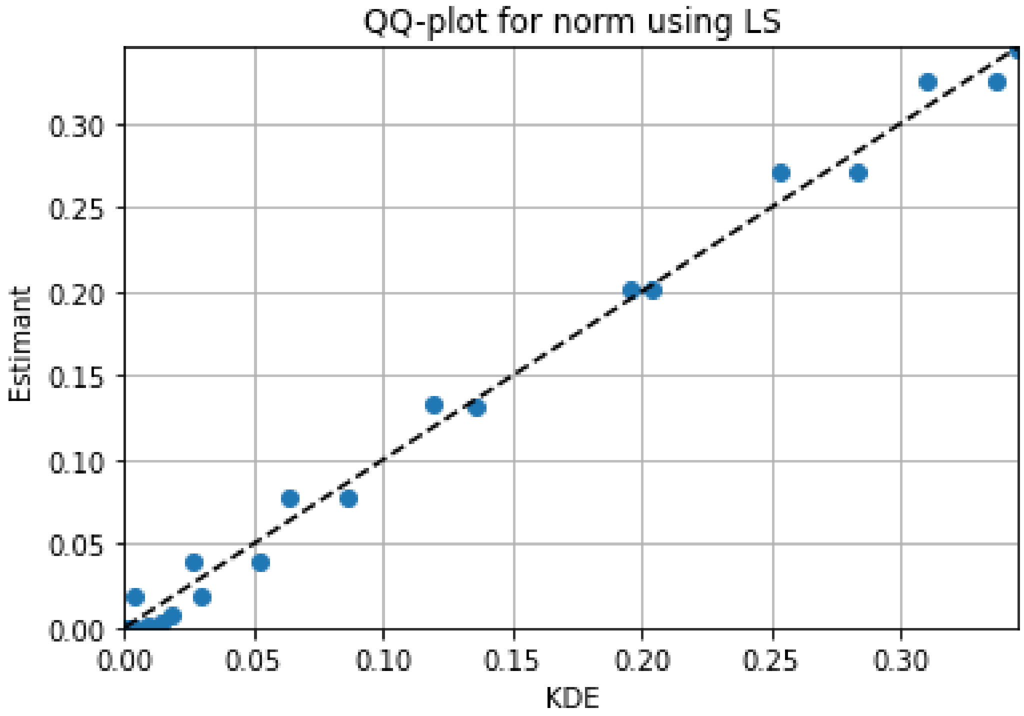
FOR lognorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.34, pvalue=0.005841778142694731), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.3978000000000015, pvalue=0.07343858790100188)

FOR exponnorm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.38, pvalue=0.0013147736033165794), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.5222000000000016, pvalue=0.03497012494432161)

FOR expon: Kolmogorov-Smirnoff test result KstestResult(statistic=0.36, pvalue=0.002834980581320342), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.7070000000000007, pvalue=0.012174301507085095)

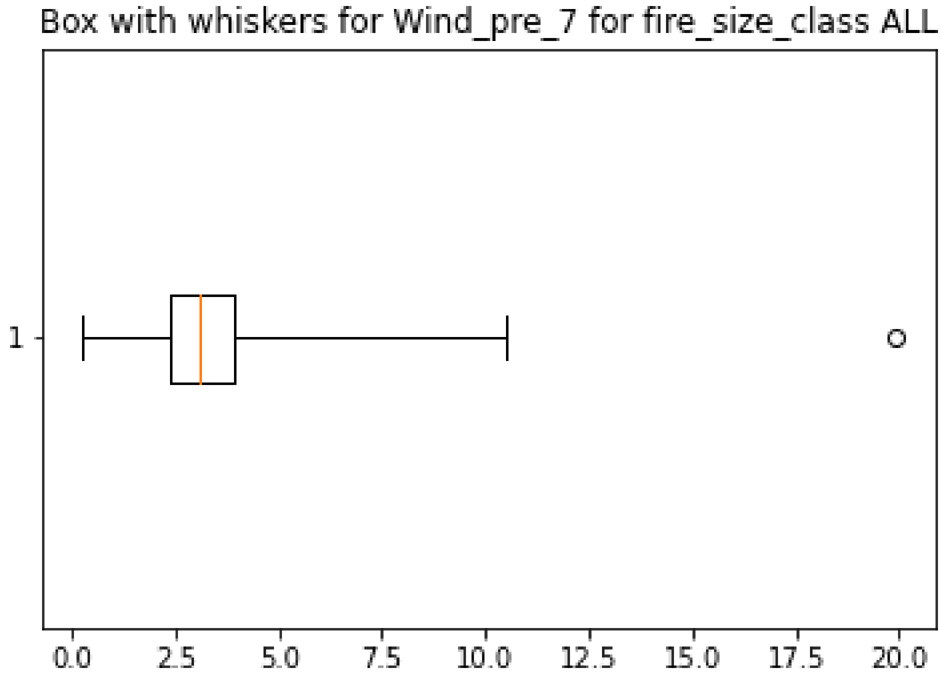
FOR norm: Kolmogorov-Smirnoff test result KstestResult(statistic=0.22, pvalue=0.17858668181221732), whereas Omega squared test (Cramér–von Mises) test CramerVonMisesResult(statistic=0.22419999999999973, pvalue=0.22729937430482294)

And according to the tests it is best in this case too.



Based on that the data for *Wind_pre_7* is **normally** distributed.

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.

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.

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. If p-value was above 0.05, we treated the distribution as possible and worth reckoning. As it is always a sample of real-world data in our case, even p-value bigger than 0 can be considered. The most important is taht it is not 0. Very often the p-value was rather high and thanks to it, we could easily say that some distributions can be considered as models and some not.

Very often the data was normally distributed. We used several variants of the normal distribution and many of them were easily applicable. Nevertheless, sometimes exponential funtion was a better approximation. Often many distributions were similarly good, and we chose just one which seemed the best.

The tests were checked between KDE and PDF of the approximant function.

Sourcecode

https://github.com/PatrykStronski/MultivariateAnalysis_Task1

(branch *main*)