Report on learning practice # 1

Analysis of univariate random variables

Performed by:

Patrik Stronski

Ivan Pavelev

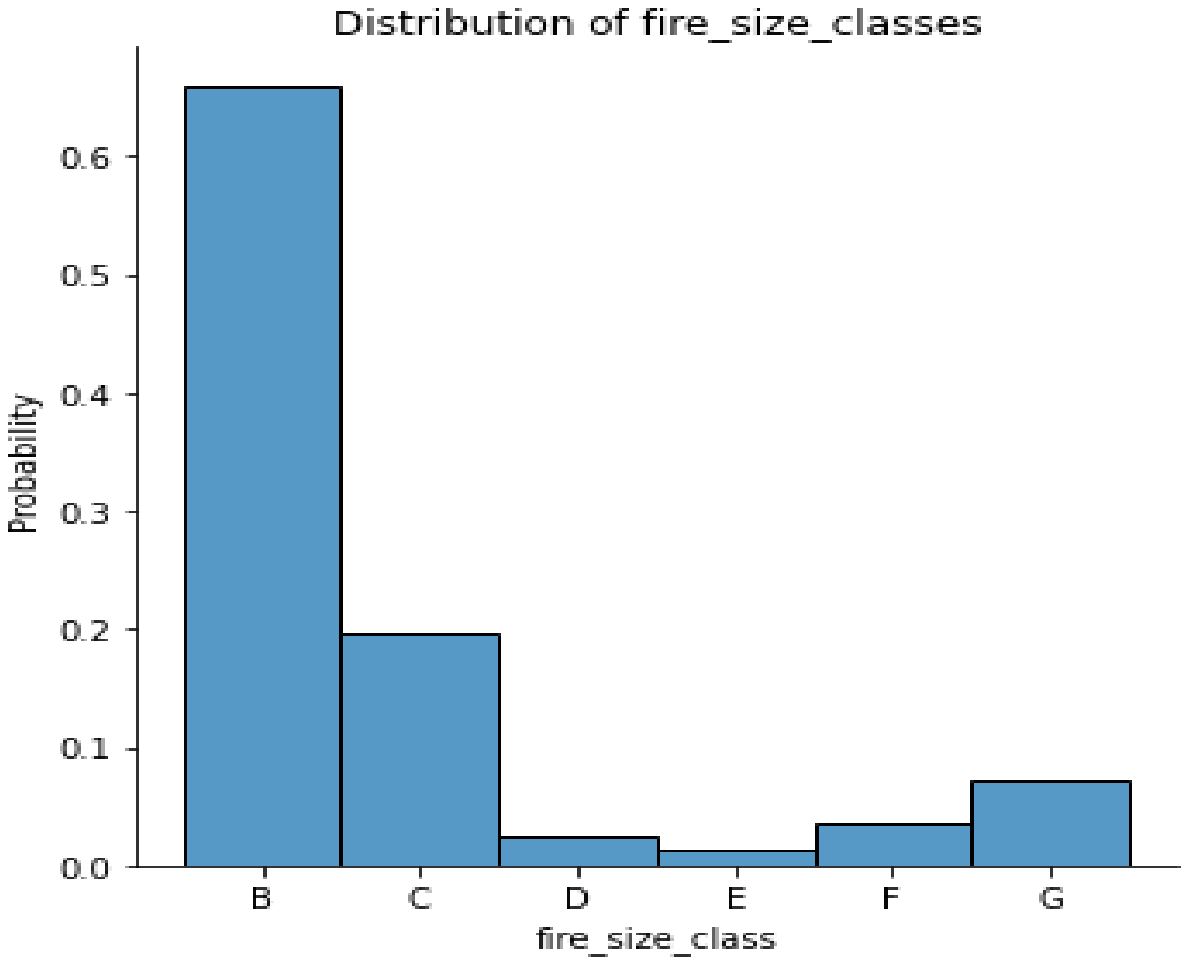
J4133C

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;

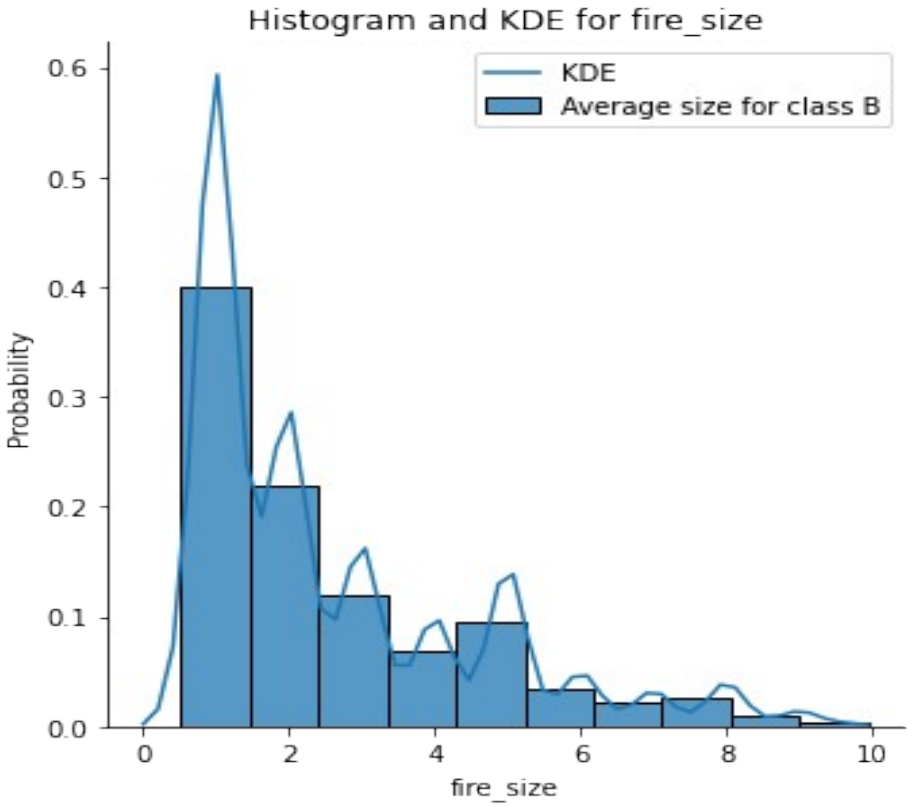
We have chosen to assess 4 variables from our dataset:

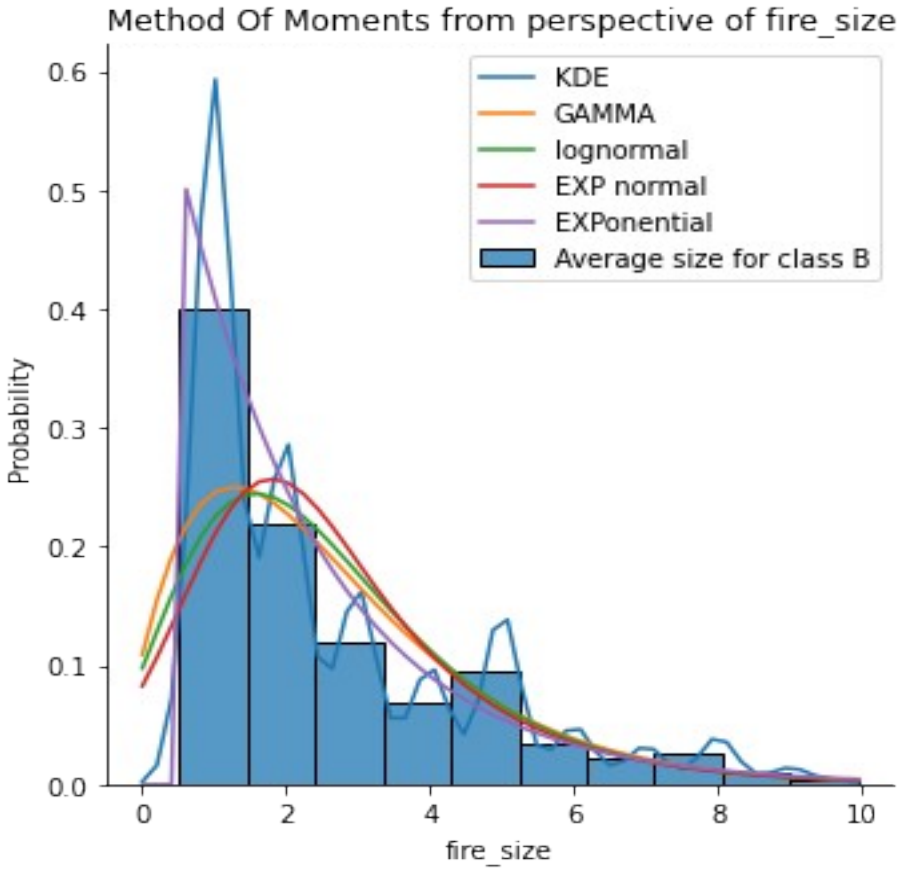
- fire_size — size of fire (in acres)
- Temp_pre_7 — temperature 7 days before the fire was discovered (in degress Celsius)
- Hum_pre_7 — humidity 7 days before the fire was discovered (in %)

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

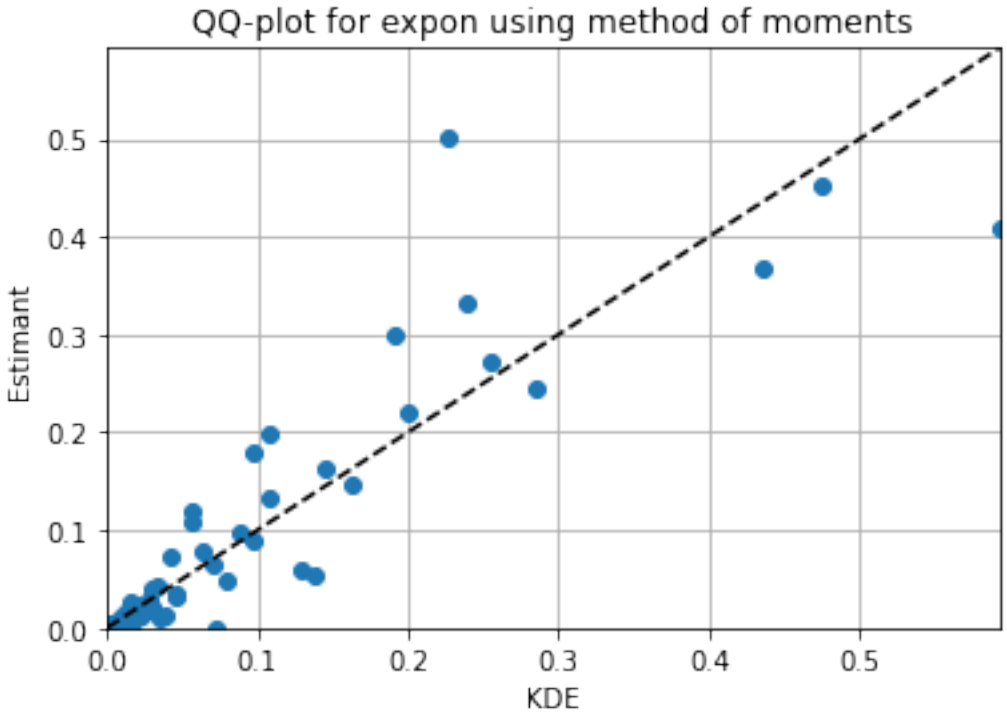
Fire_size we model for each class (how big is the fire per some class), whereas temperature, humidity, wind we model using the whole dataset.

1.1 Modelling fire_size from the perspective of the fire_size class:

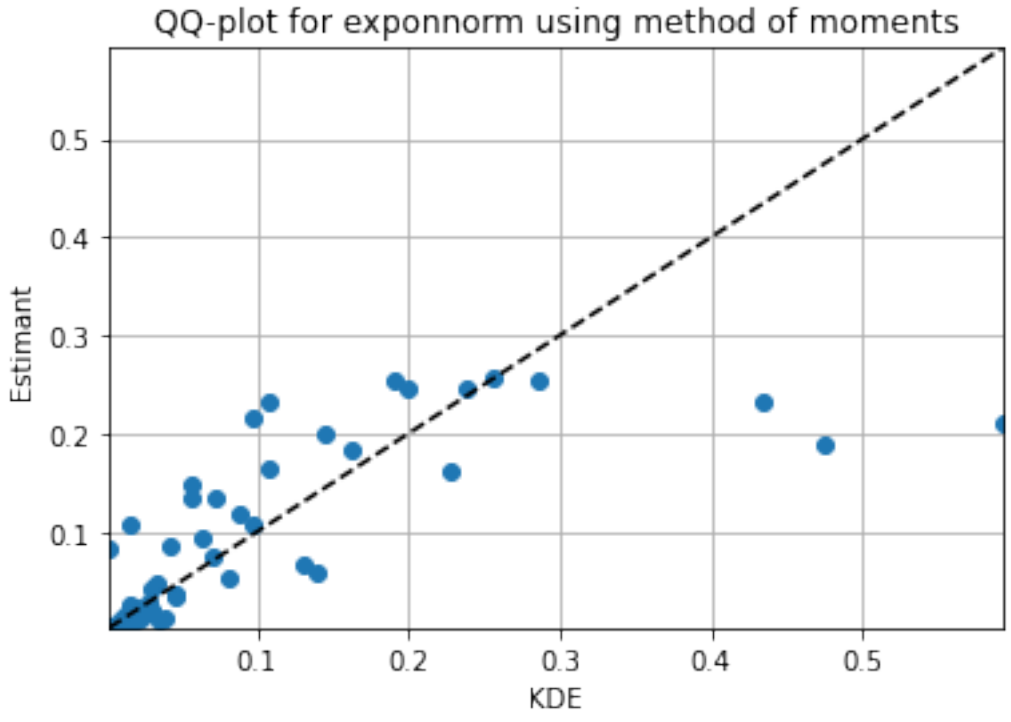


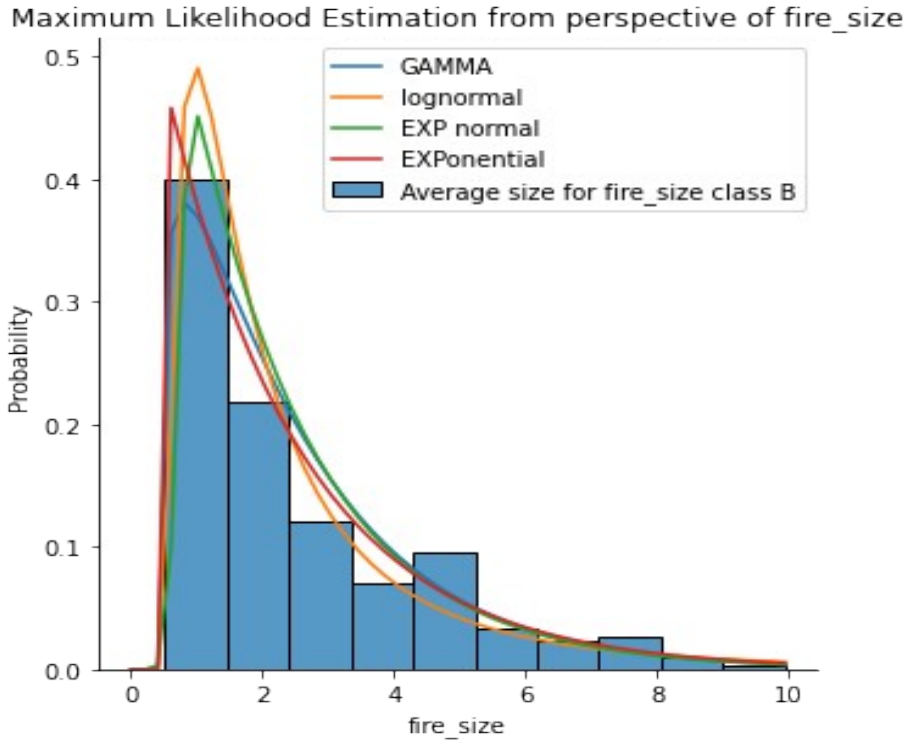


Based on the Kolmogorov-Smirnov test the best distribution chosen for modelling fire_size was done using Method of Moments for exponential (statistic=1.0, pvalue=0.0)

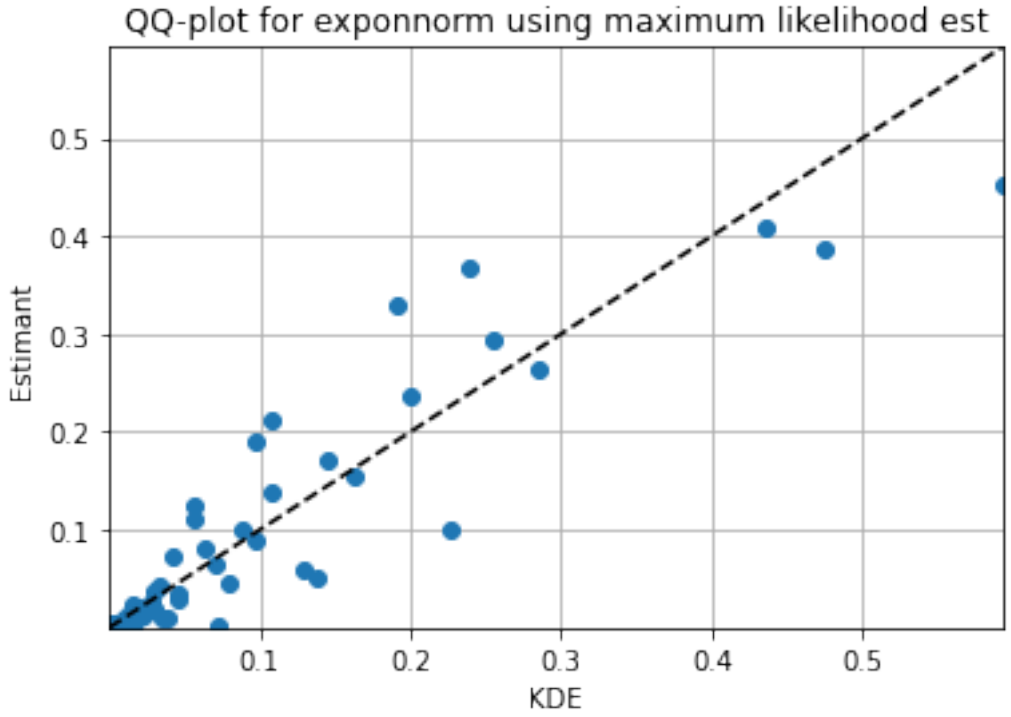


Based on the Cramer-von Mises test the best distribution chosen for modelling fire_size was done using Method of Moments for exponnorm (statistic=13.717664711218509, pvalue=3.6636860212269085e-10)

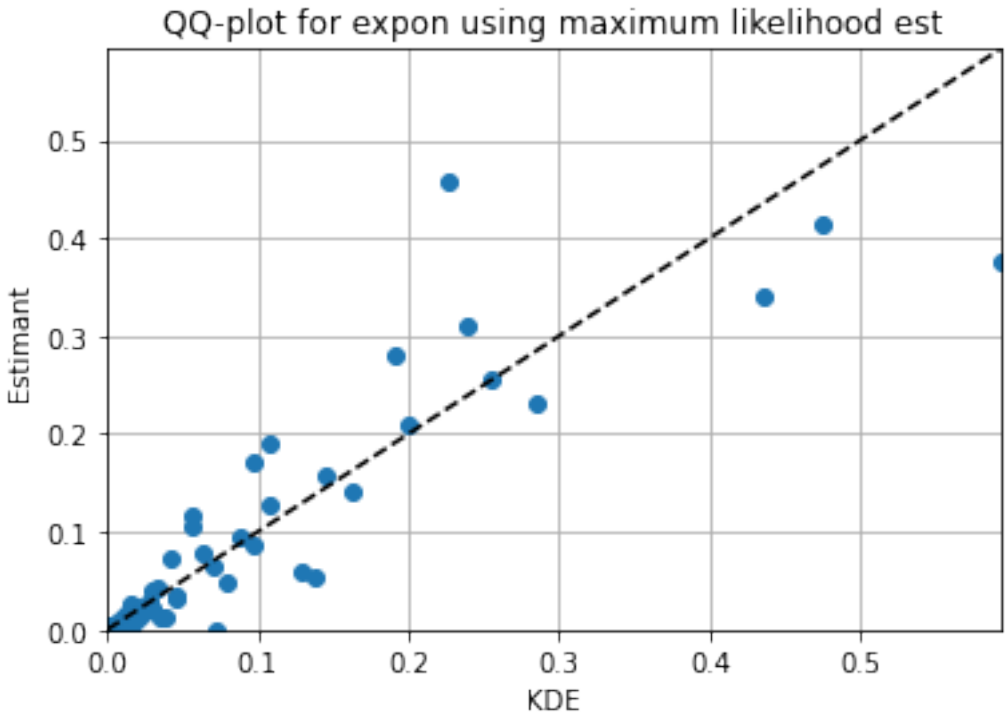


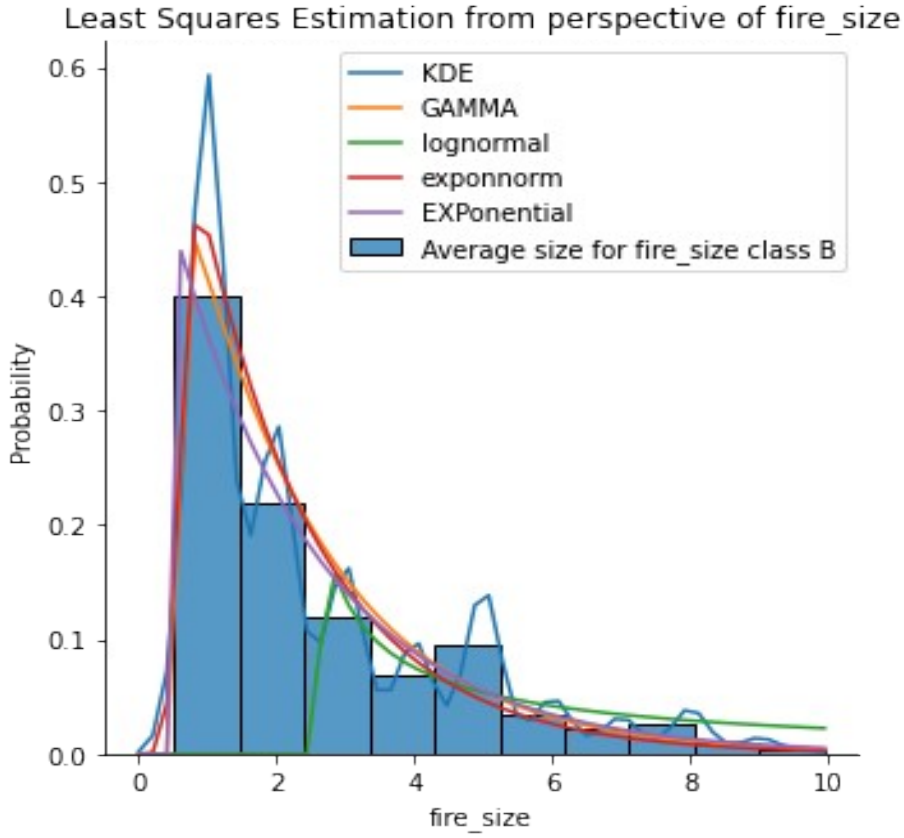


Based on the Kolmogorov-Smirnov test the best distribution chosen for modelling fire_size was done using MLE for exponnorm (statistic=0.9945510118906981, pvalue=1.3086484809461167e-113)

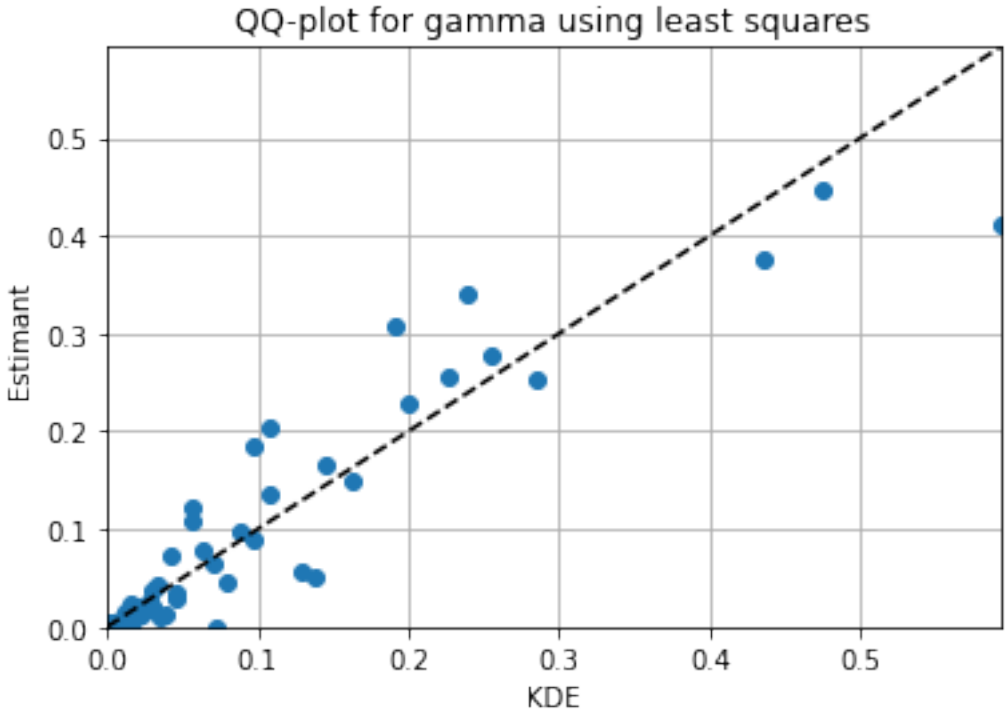


Based on the Cramer-von Mises test the best distribution chosen for modelling fire_size was done using MLE for exponential (statistic=16.59065545009577, pvalue=4.805993381040707e-10)

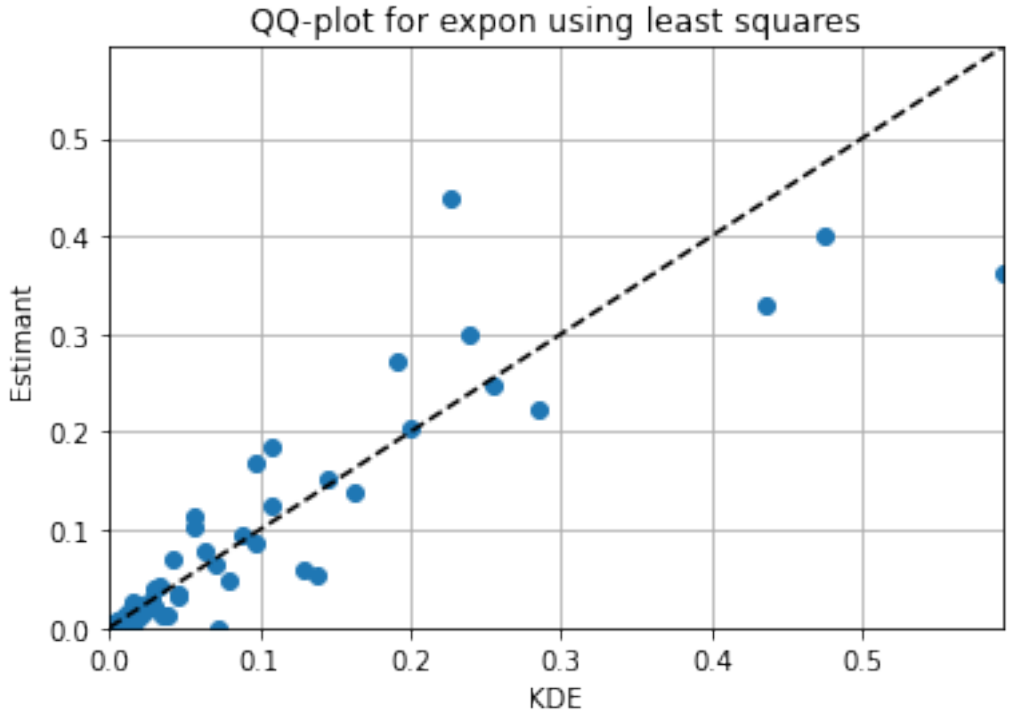




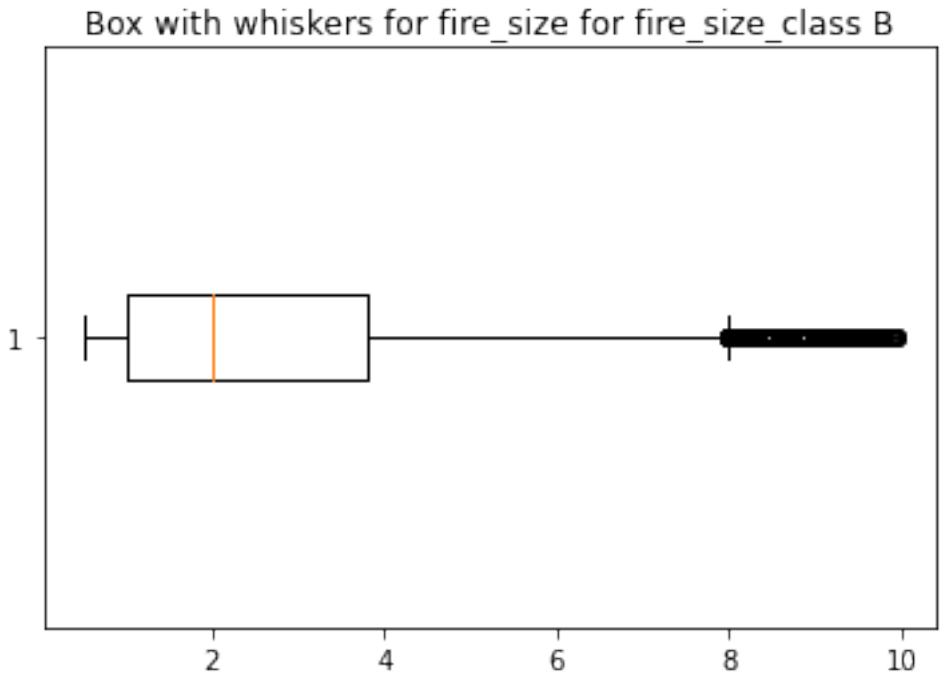
Based on the Kolmogorov-Smirnov test the best distribution chosen for modelling fire_size was done using least squares for gamma (statistic=1.0, pvalue=0.0)



Based on the Cramer-von Mises test the best distribution chosen for modelling fire_size was done using least squares for exponential (statistic=16.528761659414524, pvalue=4.490201543916328e-10)



Box with whiskers for fire_size:



Sourcecode

https://github.com/PatrykStronski/MultivariateAnalysis_Task1