Federal State Autonomous Educational Institution of Higher Education
ITMO University

ITMO UNIVERSITY

Methods and models for multivariate data analysis

Report on learning practice # 1

Analysis of univariate random variables

Performed by:
Nemat Allah Aloush
(J4134c)

Saint-Petersburg
2021

Table of Contents

Dataset

The selected dataset for the laboratory works is “House Appliances Energy Dataset”. It is an experimental data used to create regression models of appliances energy use in a low energy building.

The dataset contains 29 different features; all features are continuous except one. There are 19735 record in the dataset with no missing values. The information in the data are:

- Temperature and humidity inside different rooms in the house. (18 columns)
- Temperature, humidity, pressure, wind speed, visibility and temperature of dew point outside the house. (6 columns)
- Appliances: the energy used in the house. (1 column)
- Lights: energy use of light fixtures in the house. Which is the discreet column. (1 column)
- 2 random variables. (2 columns)
- Date and time of measuring the variables. (1 column).

The variables are in celsius, in watt, or in percentage. The data is measured each ten minutes for about 4.5 months.

The dataset can be found in the following link
https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction

1.Substantiation of chosen subsample

For this lab work, four continuous variables were chosen to investigate them. The chosen variables are:

1. T2 : The temperature in the living room.
2. RH_2 : The humidity in the living room.
3. T6 : The temperature in the outside north of the house.
4. T_out: The temperature outside the house.

2.Plotting a non-parametric estimation of PDF

First, I will list the Non-parametric estimation of PDF in form of histogram for the four variables in figure (1) and figure (2):

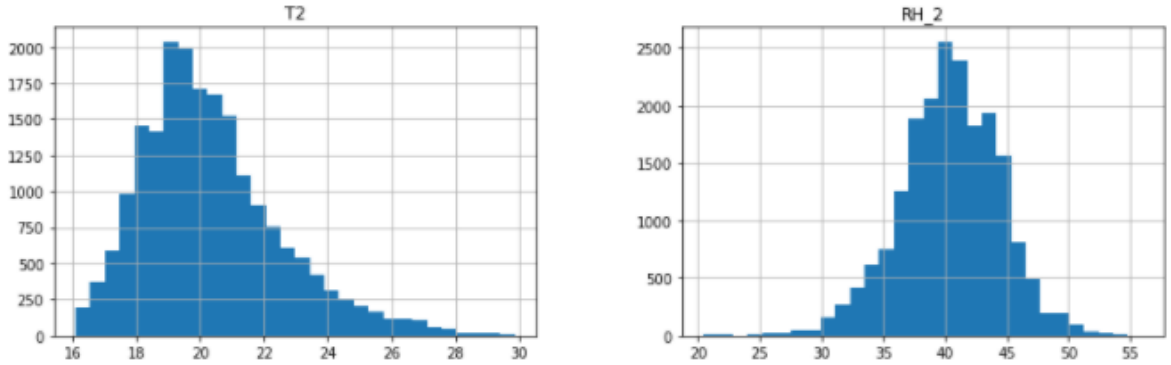

Figure 1 Non-parametric estimation of PDF in form of histogram (T2,RH_2)

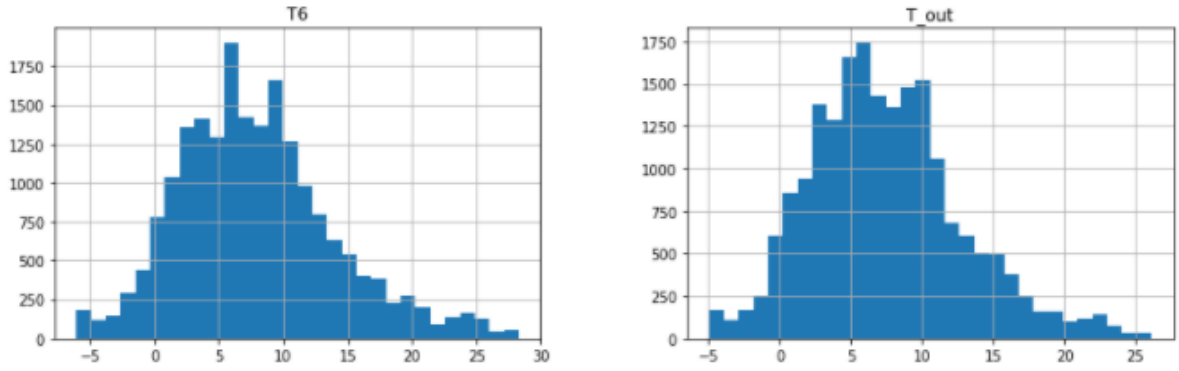

Figure 2 Non-parametric estimation of PDF in form of histogram (T6,T_out)

After observing the shape of the distribution of the data in the type of histogram, I will further investigate non-parametric estimation of PDF kernel density function for those continuous in figure (3) and figure (4).

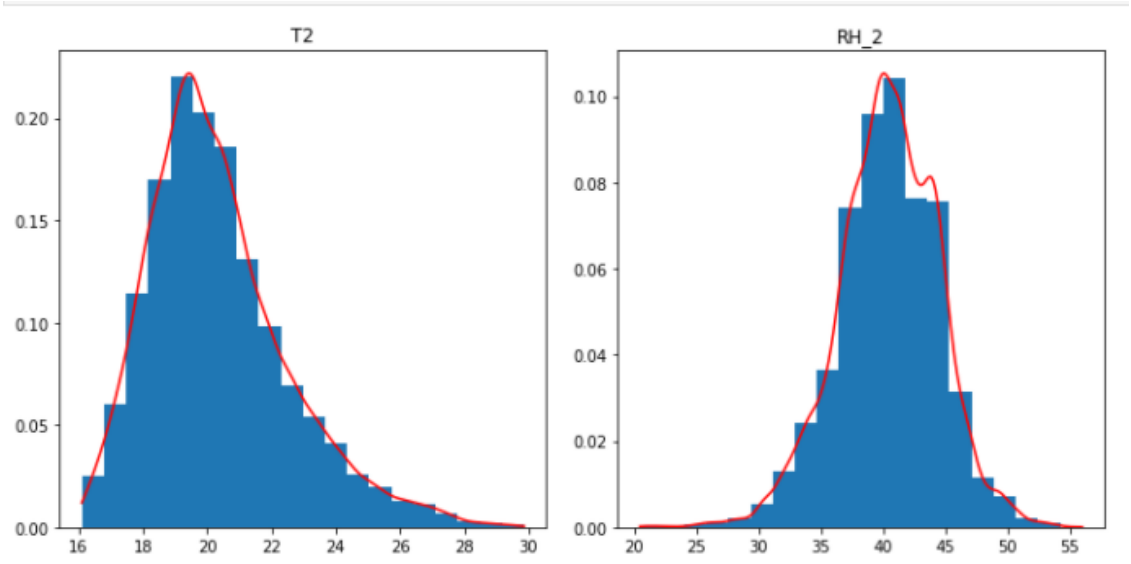

Figure 3 Non-parametric estimation of PDF kernel density function (T2, RH_2)

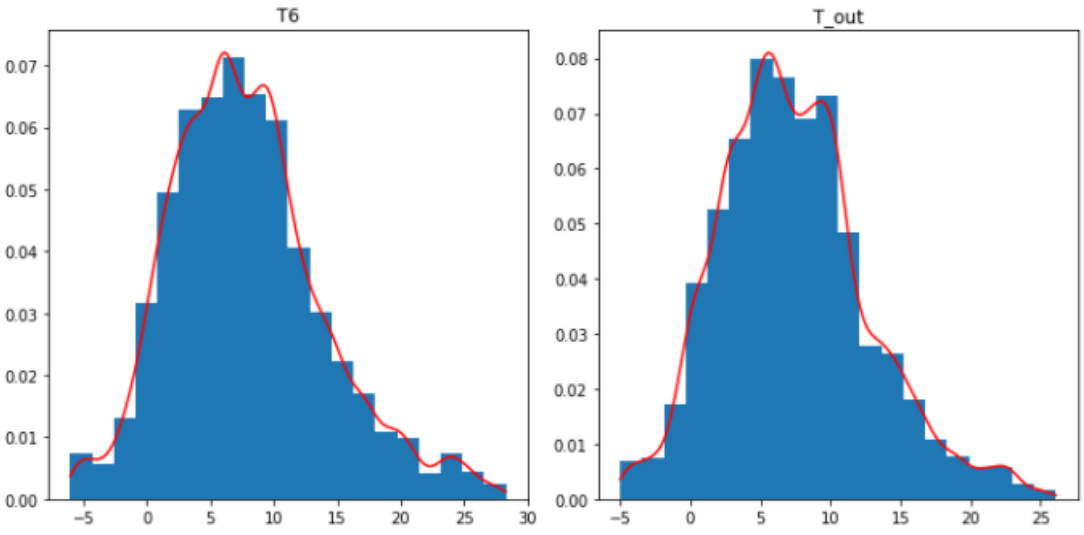

Figure 4 Non-parametric estimation of PDF kernel density function (T6,T_out)

Observing the previous plots, one can notice how kernel density function almost fit the data presented by the histograms.

3.Order statistics estimation and its representation as “box with whiskers” plot.

In the following table we can find the values of the statistics estimations for each of the objective variable.

Statistics estimation	Median	PC25	PC75	Cap bottom	Cap top
T2	20	18.79	21.5	16.1	25.56
RH_2	40.5	37.9	43.26	29.89	51.29
T6	7.3	3.63	11.25	-6.06	22.69
T_out	6.92	3.67	10.4	-5	20.4

In figure (5), we can visually observe the statistics estimations represented in box with whiskers plot, for each of the objectives variables.

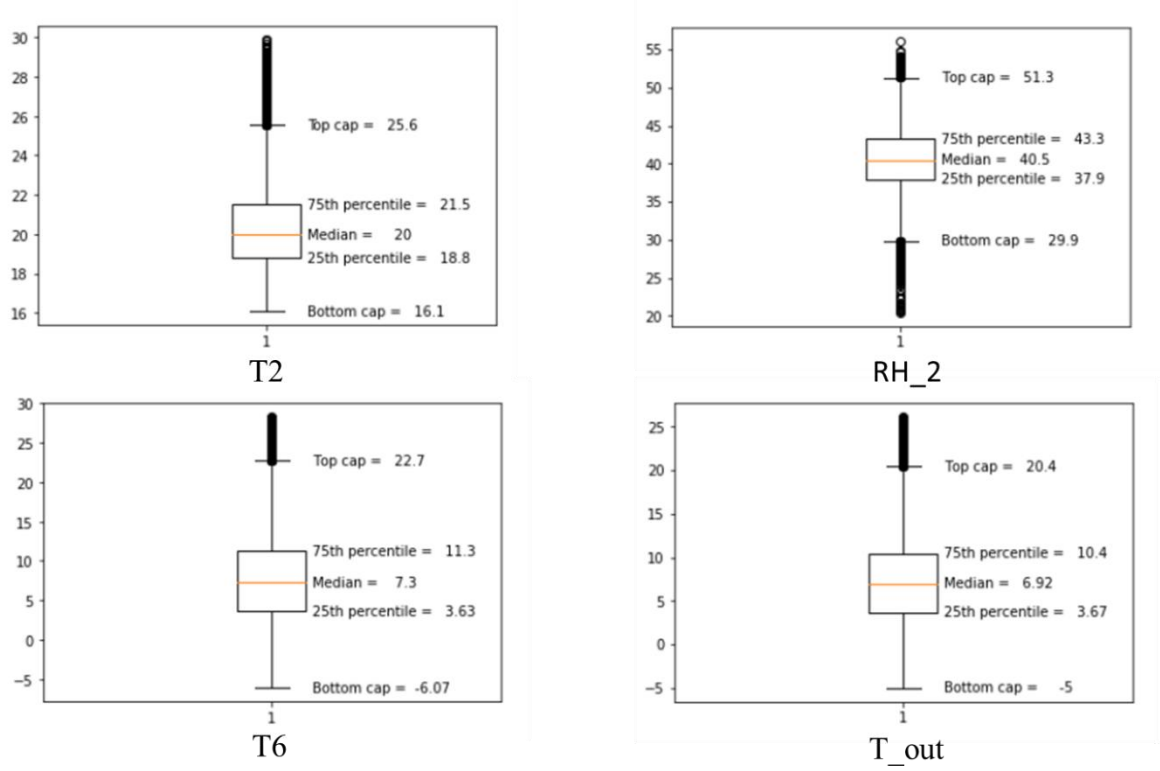

Figure 5 box with whiskers for (T2, RH_2, T6, T_out)

Through observing both numerical and visual statistics estimations, one can notice that for each of the objective random variables there is a number of outliers that are located either above the Top cap or below the Bottom Cap of the box with whiskers plot.

4.Selection of theoretical distributions that best reflect empirical data.

Theoretically, by observing the results from Non-parametric estimation of PDF kernel density function, the data looks distributed as a weighted shifted exponential distribution function. Thus, it seems that **exponentially modified Gaussian distribution** may fit the data.

Exponentially modified Gaussian distribution describes the sum of independent **normal** and **exponential** random variables. An exGaussian random variable Z may be expressed as $Z = X + Y$, where X and Y are independent, X is Gaussian with mean μ and variance σ^2, and Y is exponential of rate λ. It has a characteristic positive skew from the exponential component.

5.Estimation of random variable distribution parameters using maximum likelihood technique and LS methods.

In order to decide that a distribution fits data, we need to check p-valuse to be greater than 0.05. Unfortunately, in the case of my data none of the distributions met this condition. Many transformations were applied on the data hoping for the transformed data to fit one of the available distributions, but as a result there were no suitable distribution function for the data. The applied steps for each variable as following:

1. Try to find the best distribution function for the original data. Results in figure (6).
2. Trim the outliers and try to find the best distribution function for the trimmed data. Results in figure (7).
3. Normalize the data and trim the outliers and try to find the best distribution function for the normalized trimmed data. Results in figure (8).
4. Trying norm distribution after cox-box normalization. Results in figure (9).
5. Trying norm distribution after Yeo-Johnson normalization. Results in figure (10).

While searching for the best distribution function, the p-value was calculated by Kolmogorov–Smirnov test and compared and the distributions were found by MLE method.

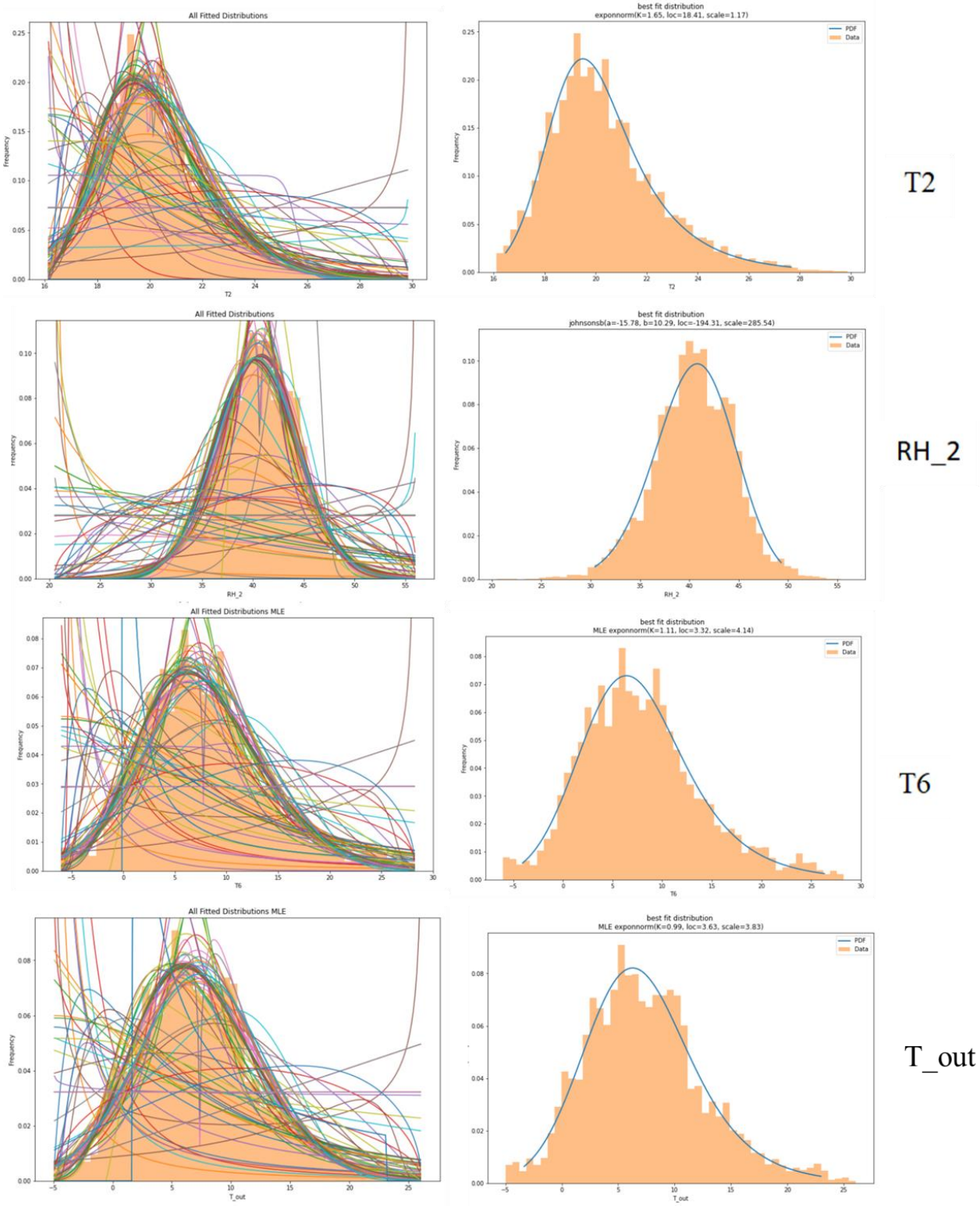

Figure 6 All and best fit distribution for original data

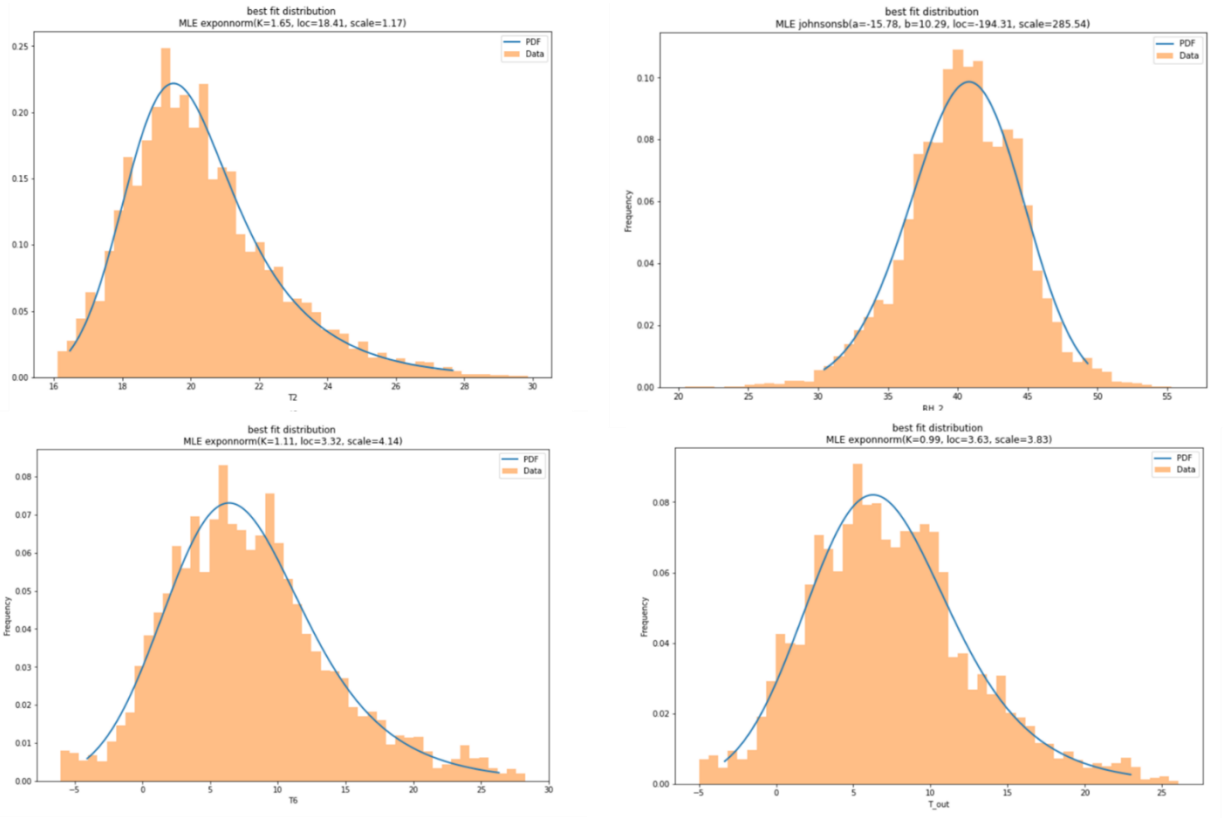

Figure 7 best distributions for trimmed data

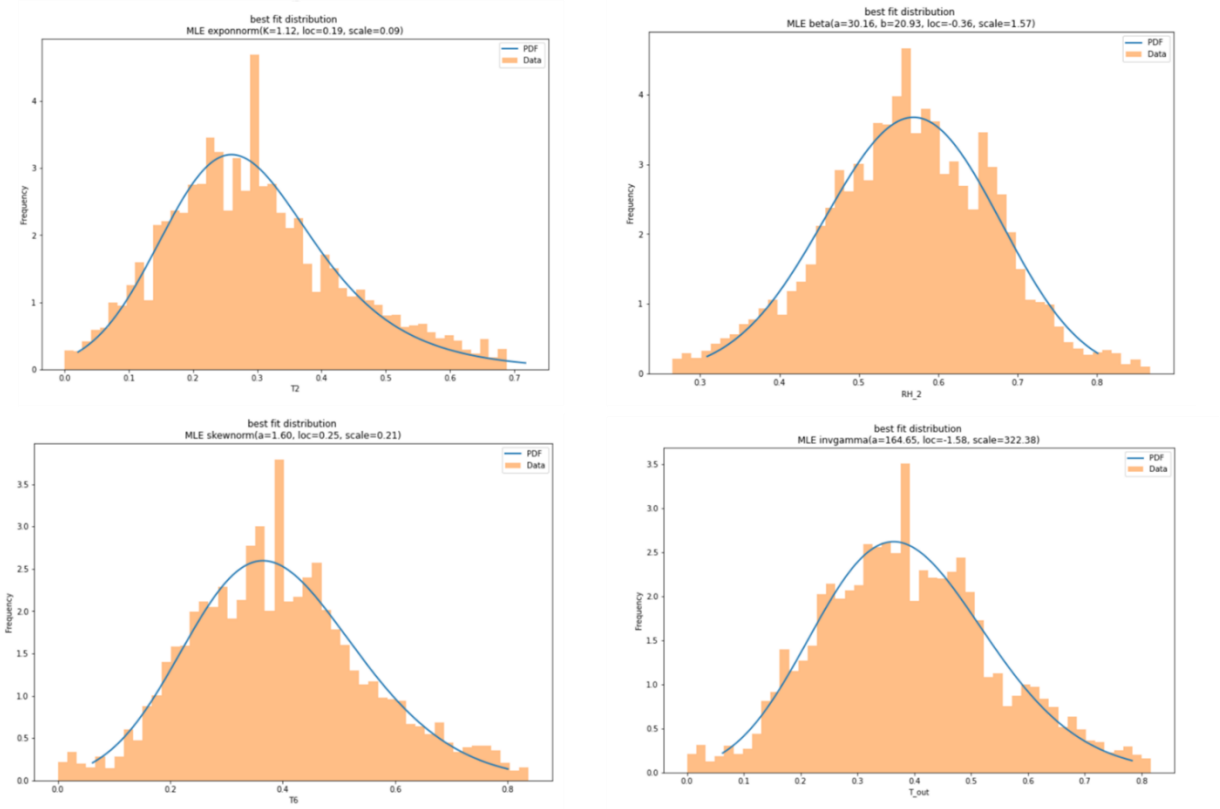

Figure 8 best distributions for normalized trimmed data

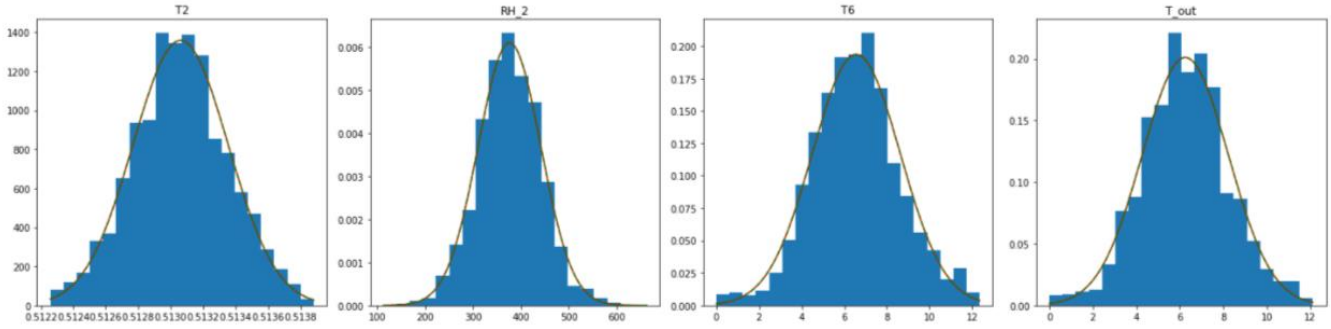

Figure 10 norm distribution after cox-box normalization

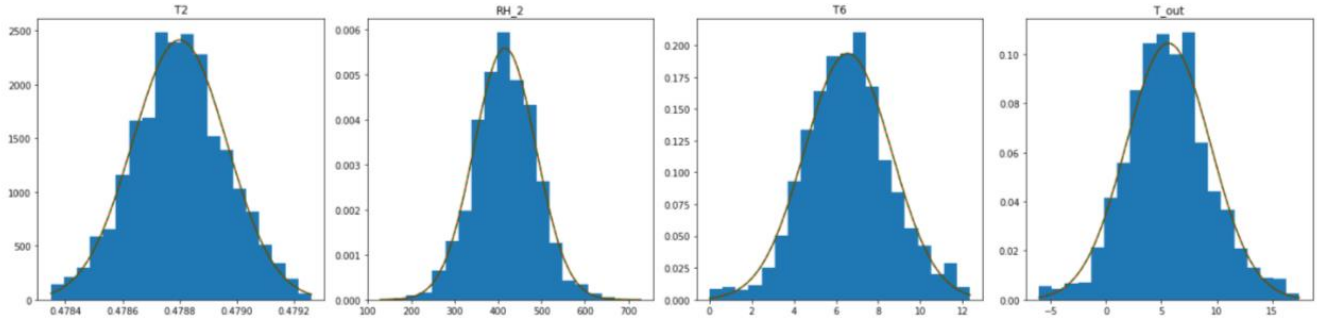

Figure 9 norm distribution after Yeo-Johnson normalization

Takinging into account that the data is real data, it is acceptable to not find a distribution that fits with p-valuse < 0.05.

After trying the previous explained steps, we can notice that for the original data the (exponnorm) distribution function was the best fit for each of the variables except for (RH_2) variable, which its best distribution fumction was (Johnsonsb). After several trials, I did not success in calculating the (Johnsonsb) distribution by, Least Square Errors methods. In addition, the p-value is not much better in (Johnsonsb) than (exponorm) for (RH_2).Thus, an (exponorm) distribution was calculated in the three methods: MLE, MM, Least Square Errors. We can see the results in the figure (10).

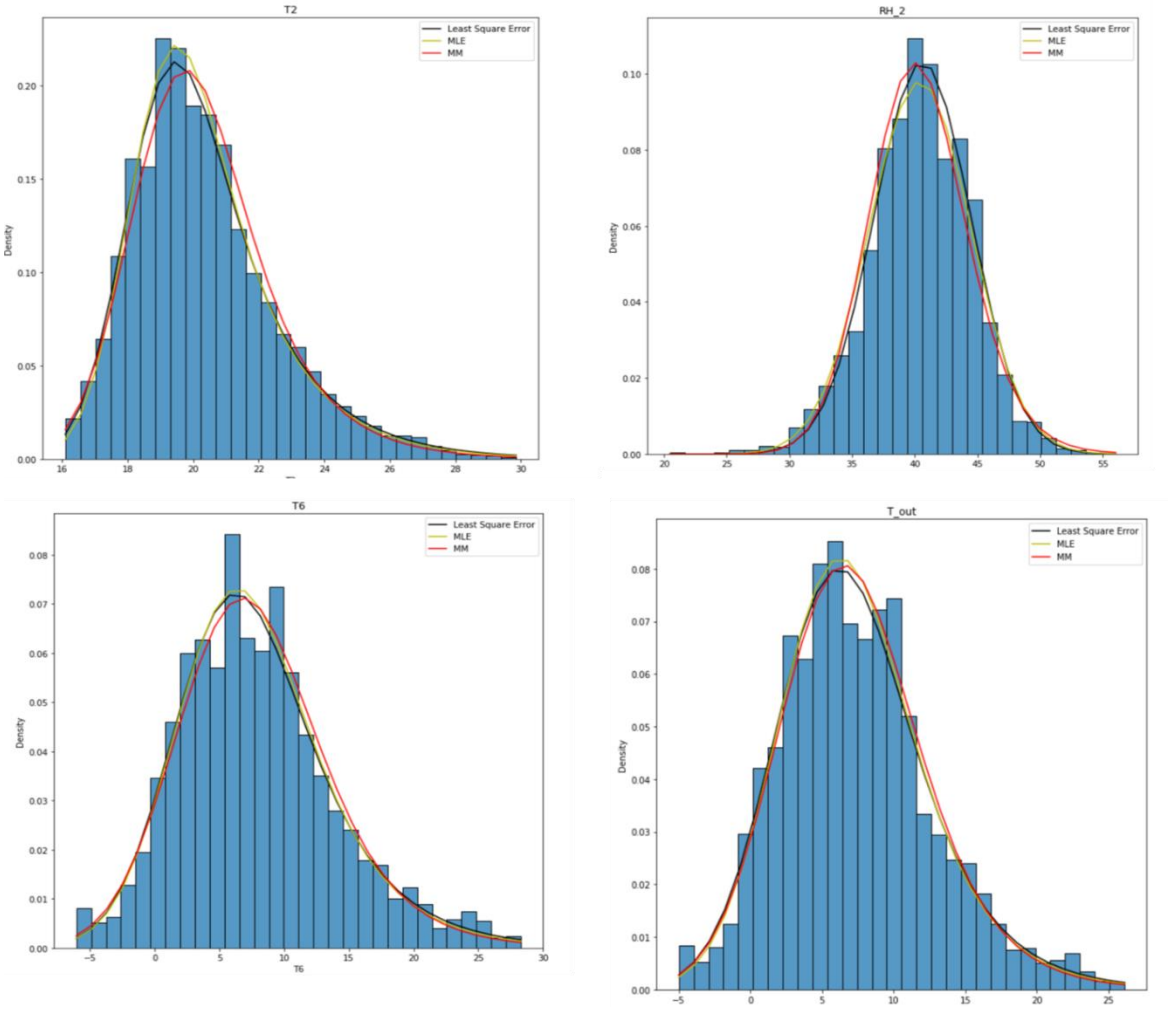

Figure 11 best distribution for each variable by 3 methods

6. Validation of empirical and theoretical distributions using quantile biplots.

In the figure (12) we can find the qq-plot for each of the variables. We can notice that there is problem with the tails in each plot, which confirm that these distributions do not fit our data enough.

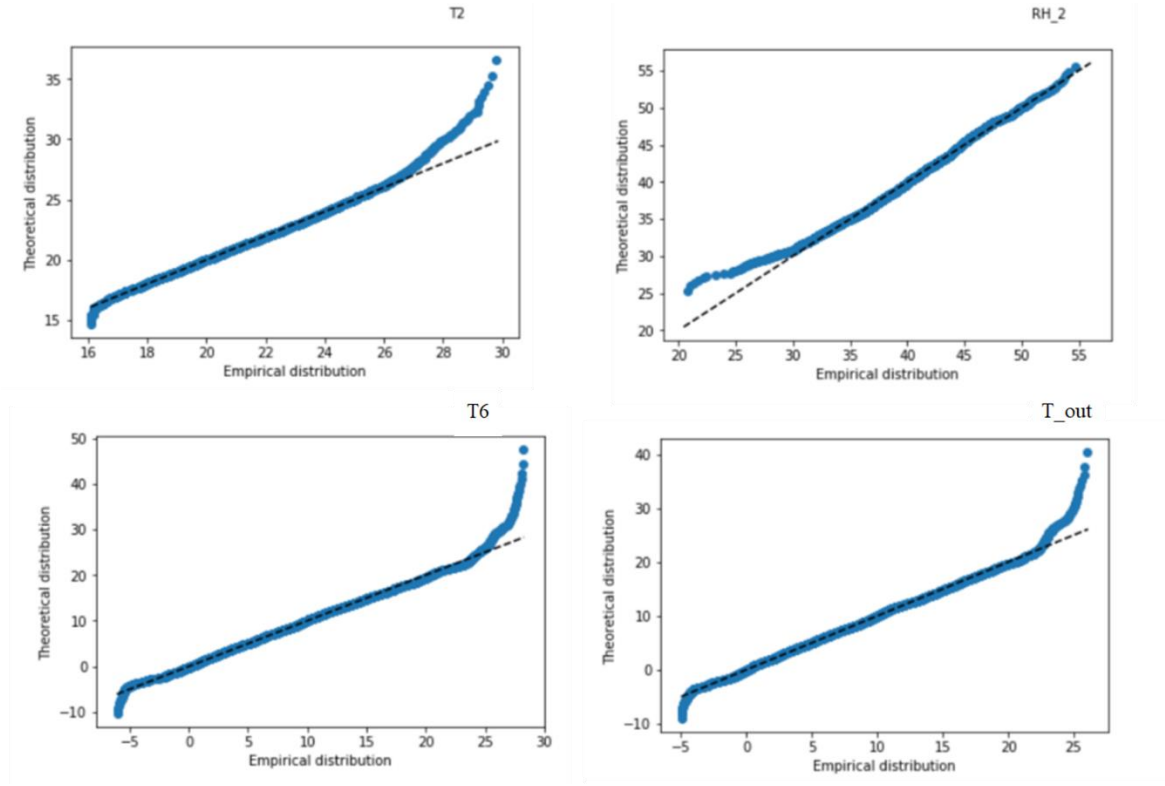

Figure 12 qq-plot for (exponnorm) distribution for eavh variable

7. Statistical tests.

In the following table, we can find the p-value calculated by two different tests for each of the methods (MLE, MM, LSE), for each variable for the (exponnorm) distribution.

We can notice how small p-value is, which is compatible with the qq-plot we got already. Since our data is real data, then it is acceptable to not find a suitable distribution function that fits with p-value > 0.05.

We can notice that in most cases, MLE seems to yield the best p-value.

	MLE		MM		LSE	
p-value	Kolmogorov–Smirnov	Cramér–von Mises	Kolmogorov–Smirnov	Cramér–von Mises	Kolmogorov–Smirnov	Cramér–von Mises
T2	0.0010040844674001886	0.016381222874696344	7.779814749044038e-23	2.2060464566209248e-10	0.0	3.916743609444495e-07
RH_2	5.080857886559252e-11	7.057592477099917e-08	4.388756596410639e-18	3.0305857823265114e-10	0.0	0.0
T6	0.00368799306166984	0.029062175469970253	9.538578646955792e-08	3.640365992330086e-05	0.0	1.9895086689203367e-07
T_out	0.0003082291457461045	0.006021171144218296	2.756377467470613e-06	0.0004345507681864369	0.0	1.7948755437480202e-07

8.Source code

Please find my code in the following github link:

https://github.com/neematAllosh/MultiderivativeDataAnalysis/blob/master/lab_01.ipynb