Federal State Autonomous Educational Institution of Higher Education

ITMO University

Methods and models for multivariate data analysis

Report on learning practice # 2

Analysis of multivariate random variables

Performed by:

Nemat Allah Aloush

(J4134c)

Saint-Petersburg

2021

Contents

Substantiation of chosen subsample

For this lab work, seven continuous variables were chosen to investigate them. The chosen variables are: (T2 , T4, T5, RH_&, RH_3, RH_out, and Appliances). Additional variable is selected. 'light' variable, which is a discrete variable, and it is not included in the analysis among the lab work. It was selected only for the conditional analysis.

1. Non-parametric estimation of PDF in form of a histogram and kernel density function.

In the figure (1), we can find the non-parametric estimation of PDF in form of a histogram and kernel density function for each of the variables.

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

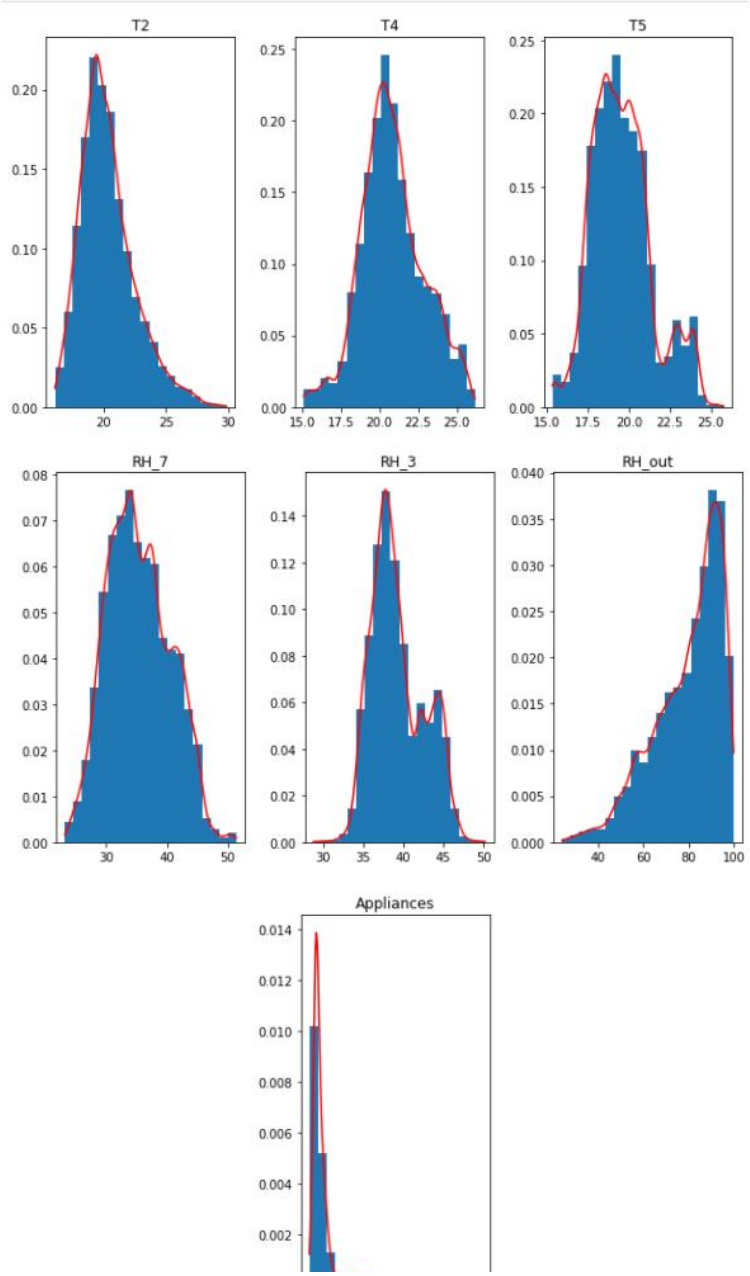

Figure 1 PDF in form of a histogram and kernel density function for all variables

Moreover, in figure (2) we can find plot of pairwise relationships in a dataset using kernel density method.

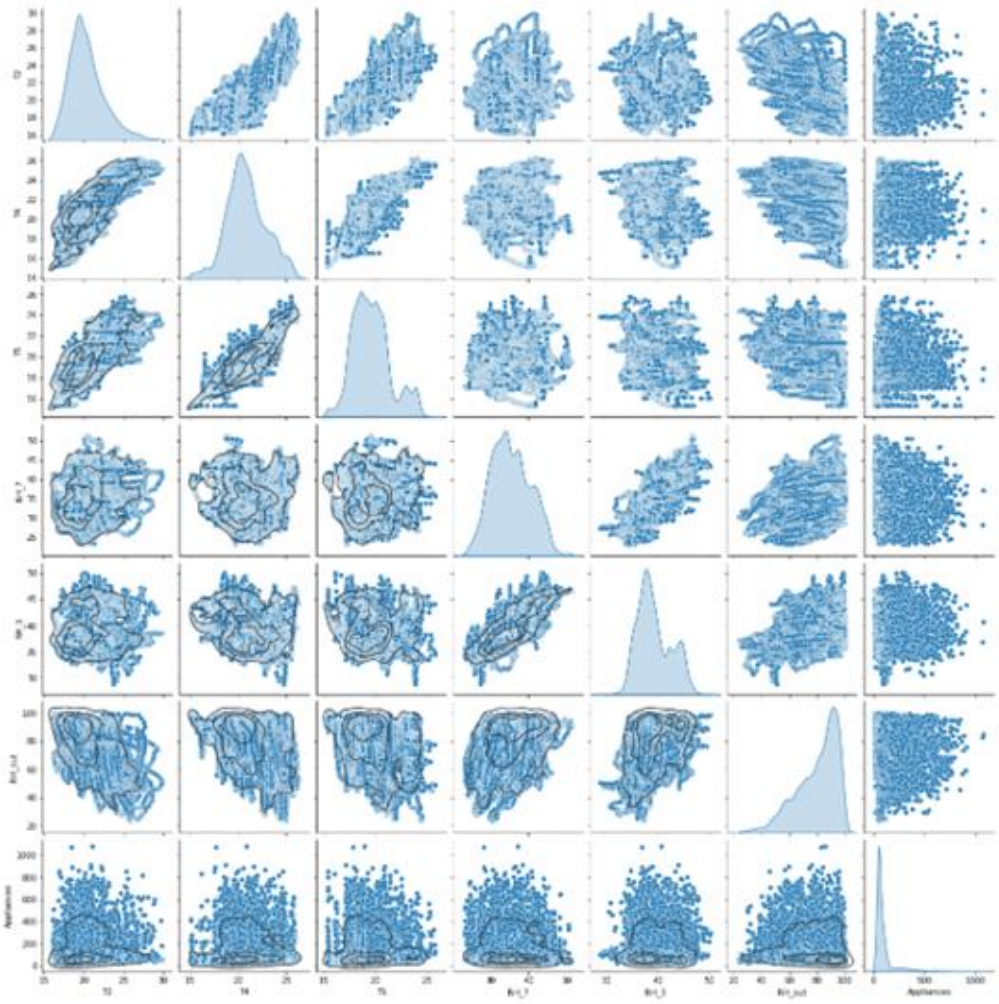

Figure 2 Multivariate Distributions

2.Estimation of multivariate mathematical expectation and variance

In the following table, we can find the mean and variance for each variable. We can notice that in the case of 'Appliances' and 'RH_out' variables, the variance is quite large and a large variance indicates that numbers in the set are far from the mean and far from each other. For other values, the variance seems smaller and that means the points are closer to the mean and to each other.

	Mean/Expectation	Variance
T2	20.341219	4.809133
T4	20.855335	4.173377
T5	19.592106	3.402635
RH_7	35.388200	26.155124
RH_3	39.242500	10.592268
RH_out	79.750418	222.042410
Appliances	97.694958	10511.353180

The following matrix is the covariance matrix, which includes the pairwise covariance of the variables.

	lights	T2	T4	T5	RH_7	RH_3	RH_out	Appliances
lights	62.979899	-0.097836	-0.143620	-1.152735	1.423334	3.387651	8.105521	160.512176
T2	-0.097836	4.809133	3.414049	2.914777	2.570685	0.867146	-16.511737	26.996601
T4	-0.143620	3.414049	4.173377	3.285300	0.454756	-0.933861	-11.829524	8.436711
T5	-1.152735	2.914777	3.285300	3.402635	1.404742	-0.300548	-7.530112	3.736981
RH_7	1.423334	2.570685	0.454756	1.404742	26.155124	13.859693	28.845903	-29.175191
RH_3	3.387651	0.867146	-0.933861	-0.300548	13.859693	10.592268	17.274146	12.109644
RH_out	8.105521	-16.511737	-11.829524	-7.530112	28.845903	17.274146	222.042410	-232.646723
Appliances	160.512176	26.996601	8.436711	3.736981	-29.175191	12.109644	-232.646723	10511.353180

3.Non-parametric estimation of conditional distributions, mathematical expectations and variances.

- In the following table, we can see the conditional mean for the variables, where the condition is for the 'light' district value to take a fix value. The values of 'light' variables are (0,10,20,30,40,50,60,70).

Conditional mean

light	T2	T4	T5	RH_7	RH_3	RH_out	Appliances
0	20.3632	20.8968	19.6805	35.3249	39.045	79.2606	86.5847
10	20.0916	20.3328	19.1825	35.4423	39.6249	80.7906	129.037
20	20.4986	21.2078	19.5228	35.5511	39.9082	81.3914	136.429
30	20.3294	20.9572	19.2241	35.8167	40.5394	82.8748	150.215
40	19.9801	19.7088	18.2575	39.211	43.3686	88.8766	182.338
50	20.1456	19.4059	17.751	38.9325	42.7615	83.4074	178.889
60	19.4267	19	17.1	44.2633	44.8267	91.1667	580
70	19.3567	18.89	17.1	42.7175	44.9	91.3333	230

- In the following table, we can see the conditional variance for each of the variables, where the condition is for the 'light' district value to take a fix value at each time. The values of 'light' variables are (0,10,20,30,40,50,60,70). The values of variances for values of light (60, 70) are zeros because there is only one row in the dataset that has a value of light equal to 60 and only one row in the dataset that has a value of light equal to 70, as shown in the conditional distributions in figure (6) for light value equal to 60, and figure (7) equal to 70.

Conditional variance

light	T2	T4	T5	RH_7	RH_3	RH_out	Appliances
0	5.20963	4.44548	3.5778	24.1418	9.84228	234.998	8107.97
10	3.6833	3.48059	2.85011	32.0365	13.1172	189.015	16249.8
20	3.35164	2.81608	2.55222	32.7139	11.1487	166.787	16292.4
30	2.6417	2.09297	2.17262	34.69	12.188	142.259	19723
40	2.20928	1.66908	1.983	30.7499	10.1088	60.672	16277.7
50	1.22948	2.08942	0.948241	66.5511	15.2679	195.266	21254.3
60	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0

- In the following three figures, we can see the conditional distribution of the data according to five different values of the variable (light).
  In figure (3), the conditional distribution of the data according to the value of light variable equal to Zero.

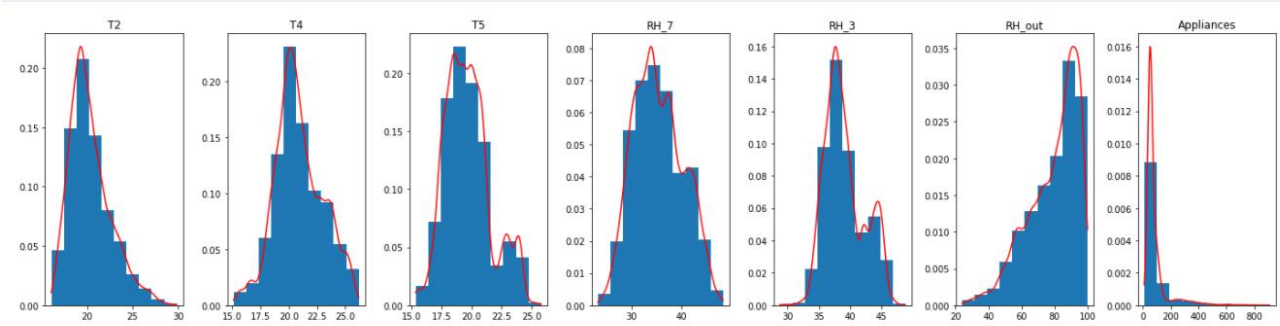

Figure 3 Conditional Distribution, light = 0

In figure (4), the conditional distribution of the data according to the value of light variable equal to Twenty.

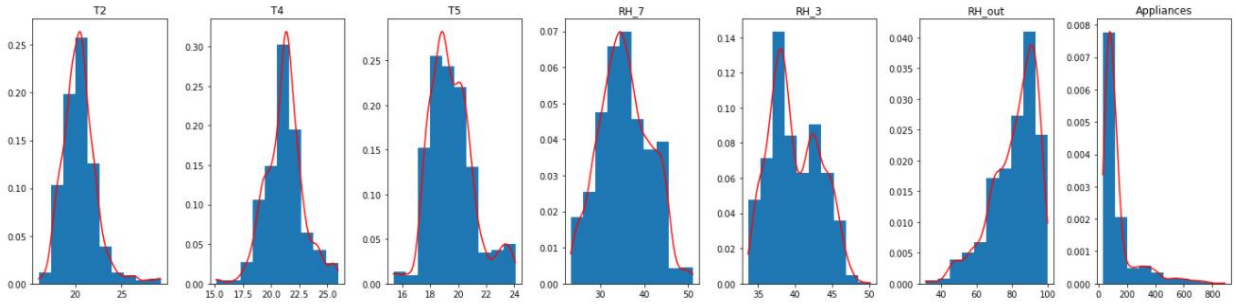

Figure 4 Conditional Distribution, light = 20

In figure (5), the conditional distribution of the data according to the value of light variable equal to Thirty.

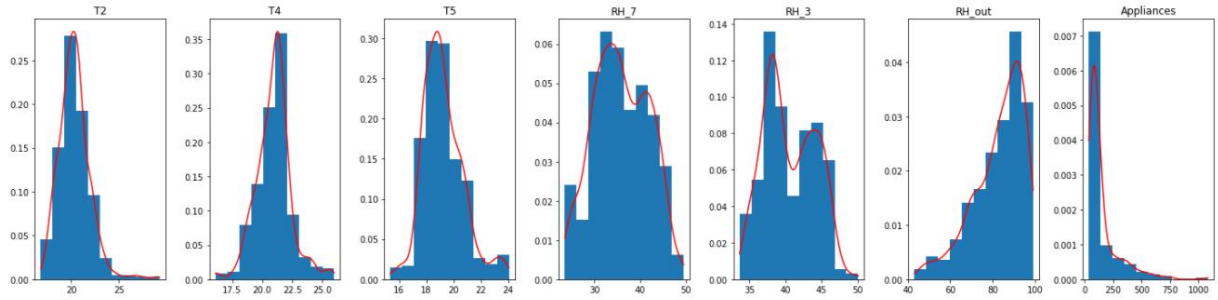

Figure 5 Figure 4 Conditional Distribution, light = 30

In figure (6), the conditional distribution of the data according to the value of light variable equal to Sixty. We can see that there is only one row in the data that has a light value equal to sixty.

	lights	T2	T4	T5	RH_7	RH_3	RH_out	Appliances
11	60	19.426667	19.0	17.1	44.263333	44.826667	91.166667	580

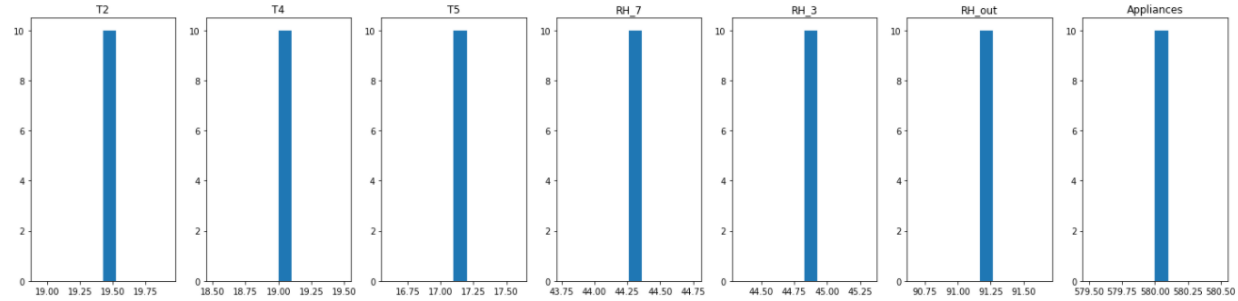

Figure 6 Conditional Distribution, light = 60

In figure (7), the conditional distribution of the data according to the value of light variable equal to Sixty. We can see that there is only one row in the data that has a light value equal to seventy.

	lights	T2	T4	T5	RH_7	RH_3	RH_out	Appliances
10	70	19.356667	18.89	17.1	42.7175	44.9	91.333333	230

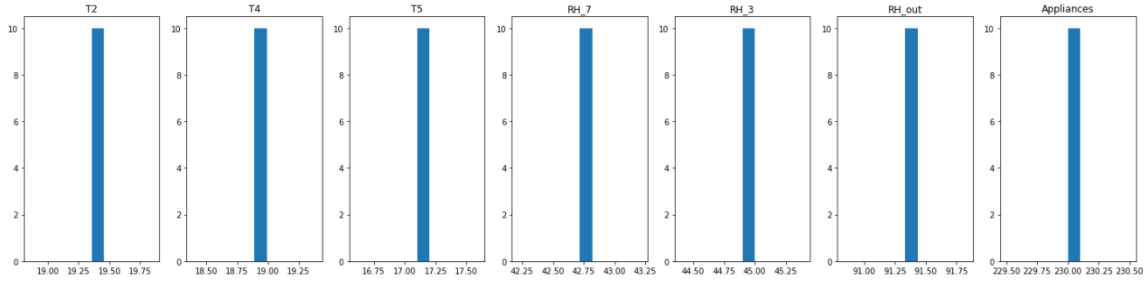

Figure 7 Conditional Distribution, light = 70

4.Estimation of pair correlation coefficients, confidence intervals for them and significance levels.

In the following figure(8), we can see the correlation matrix, we can notice that some variables has a strong correlation, for example: T2 and T4. T5 and T4. RH_3 and RH_7. T5 and Appliances

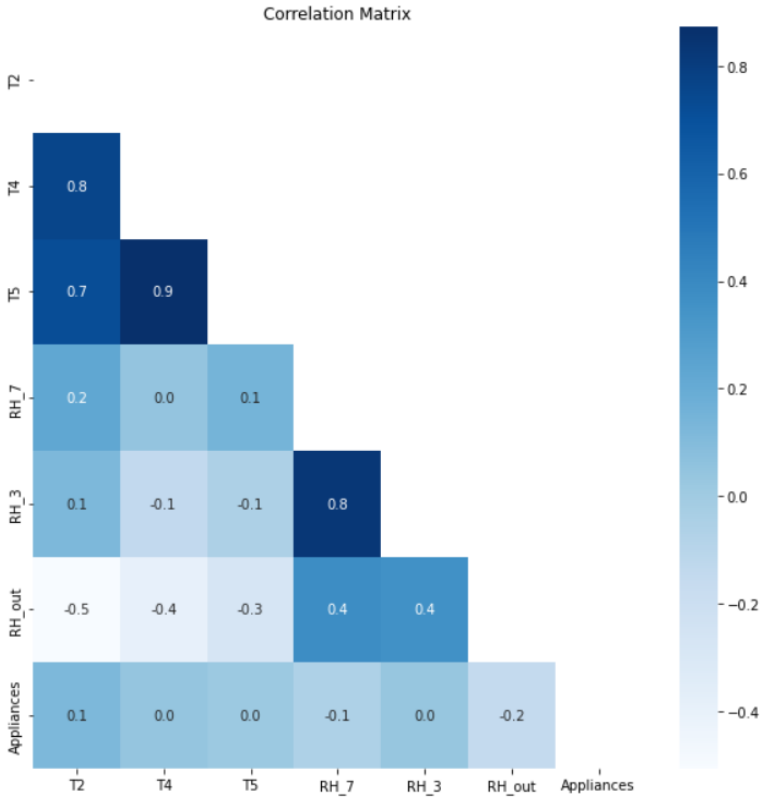

Figure 8 Correlation matrix

For the confidence intervals for correlations and significance levels, we can find this information in the following table. From the values in the table, we can notice how the significance level is quite low for some pair correlations.

First Col	Second Col	Correlation Coefficient	Confidence Interval	P: Significance Level
T2	T4	0.762066	[0.756153764670642, 0.7678539314101446]	0
T2	T5	0.72055	[0.7137738735811071, 0.7271916147913446]	0
T2	RH_7	0.229212	[0.2159505649811892, 0.24238868382341944]	1.38129e-233
T2	RH_3	0.121497	[0.10772734605041417, 0.1352193881350561]	8.80129e-66
T2	RH_out	-0.505291	[-0.5156078326220358, -0.4948273171448083]	0
T2	Appliances	0.120073	[0.10629942931743509, 0.13380106425190136]	2.78495e-64
T4	T5	0.871813	[0.8684246130440574, 0.8751208921304521]	0
T4	RH_7	0.0435268	[0.02959282244307497, 0.05744383090477239]	9.5227e-10
T4	RH_3	-0.140457	[-0.1541071714489836, -0.12675369607921233]	1.68224e-87
T4	RH_out	-0.388602	[-0.40038355878048815, -0.37669281171867663]	0
T4	Appliances	0.040281	[0.02634385603467604, 0.054202453524728925]	1.50788e-08
T5	RH_7	0.148905	[0.13523443271282115, 0.16251970738019894]	3.1733e-98
T5	RH_3	-0.0500625	[-0.06396972382064933, -0.03613578001563801]	1.96692e-12
T5	RH_out	-0.273953	[-0.28680852818653324, -0.2609984760571263]	0
T5	Appliances	0.0197599	[0.005809530060511341, 0.033702501376384725]	0.00550345
RH_7	RH_3	0.832685	[0.8283565451364288, 0.8369140212167392]	0
RH_7	RH_out	0.378519	[0.36650267616803756, 0.39040923378286996]	0
RH_7	Appliances	-0.0556424	[-0.06954036496703897, -0.041722876444624454]	5.1873e-15
RH_3	RH_out	0.356192	[0.3439493209921373, 0.36831354778941333]	0
RH_3	Appliances	0.0362917	[0.02235110624614817, 0.05021822583534982]	3.40254e-07
RH_out	Appliances	-0.152282	[-0.1658818748888154, -0.13862497601115528]	1.07752e-102

5. Task formulation for regression, multivariate correlation.

The target value is 'Appliances'. In the following we can see the results of linear regression:

model bias: [-75.83768726]

MAE = 58.73130085560194

MSE = 10781.963099098515

MAPE Appliances = 70.01508

Determination coefficient = 0.053556956840037584

We can notice how the error values are quite large.

In figure (4), we can find the original and predicted data from the linear regression model.

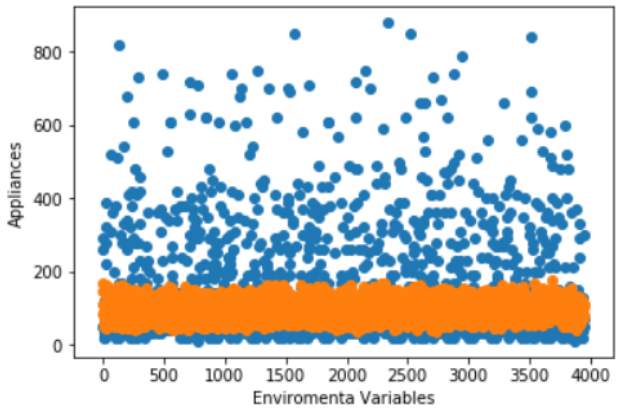

Figure 9 Linear Regression

6. Regression model,multicollinearity and regularization (if needed).

I have investigated both Lasso and Ridge models for different parameters.

- Lasso:

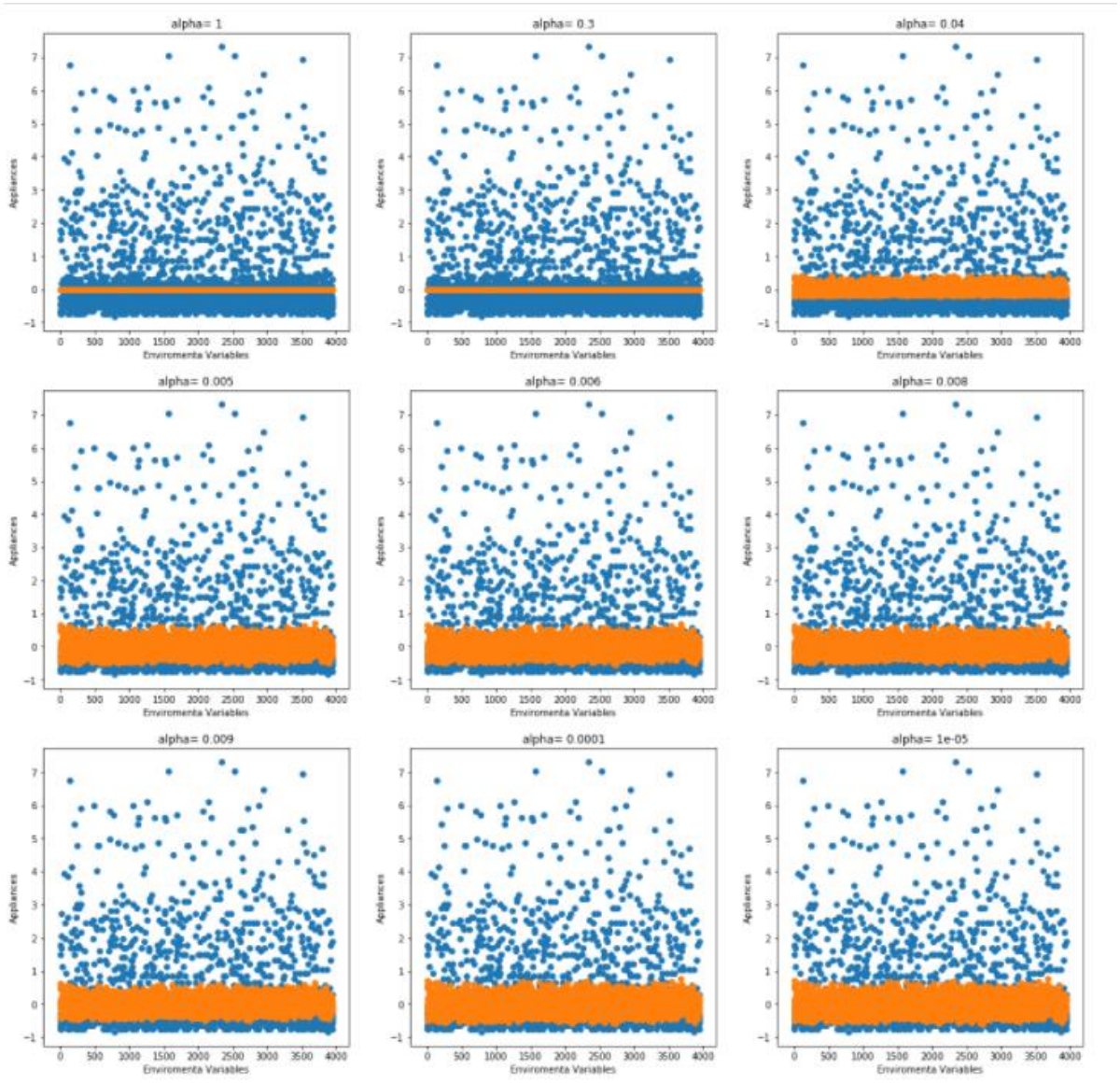

Figure 10 Lasso Regression with different alpha values

Alpha	MAE	MSE	MAPE	Cofficient	Bias
1	0.586256	1	100	0	6.52576e-18
0.3	0.586256	1	100	0	6.52576e-18
0.04	0.571321	0.968148	159.696	0.0318523	2.84648e-17
0.005	0.558181	0.945921	216.13	0.0540785	1.41359e-16
0.006	0.558189	0.946002	213.917	0.0539978	1.35219e-16
0.008	0.558312	0.946275	209.572	0.0537246	1.2294e-16
0.009	0.558441	0.946468	207.459	0.0535317	1.16824e-16
0.0001	0.558669	0.946068	227.492	0.0539317	1.71523e-16
1e-05	0.558689	0.946079	227.71	0.0539205	1.72073e-16

Looking at the plots and numerical results, we can notice how the error decrease from the linear regression and it seems like alpha = 0.04 may yield the best results.

- Ridge:

Alpha	MAE	MSE	MAPE	Cofficient	Bias
0.1	0.558691	0.946081	227.73	0.0539194	1.72125e-16
10	0.558657	0.946071	227.377	0.0539286	1.71346e-16
500	0.558096	0.946451	213.117	0.0535489	1.40696e-16
1000	0.558608	0.947648	202.939	0.0523522	1.19748e-16
10000	0.568826	0.965027	154.317	0.0349732	3.10954e-17
80000	0.580581	0.988181	114.873	0.0118195	1.68769e-18
100000	0.581456	0.989946	111.959	0.0100539	1.70534e-18
200000	0.583487	0.994203	105.119	0.00579707	2.84788e-18
1e+06	0.585531	0.998665	99.3349	0.00133489	5.49397e-18

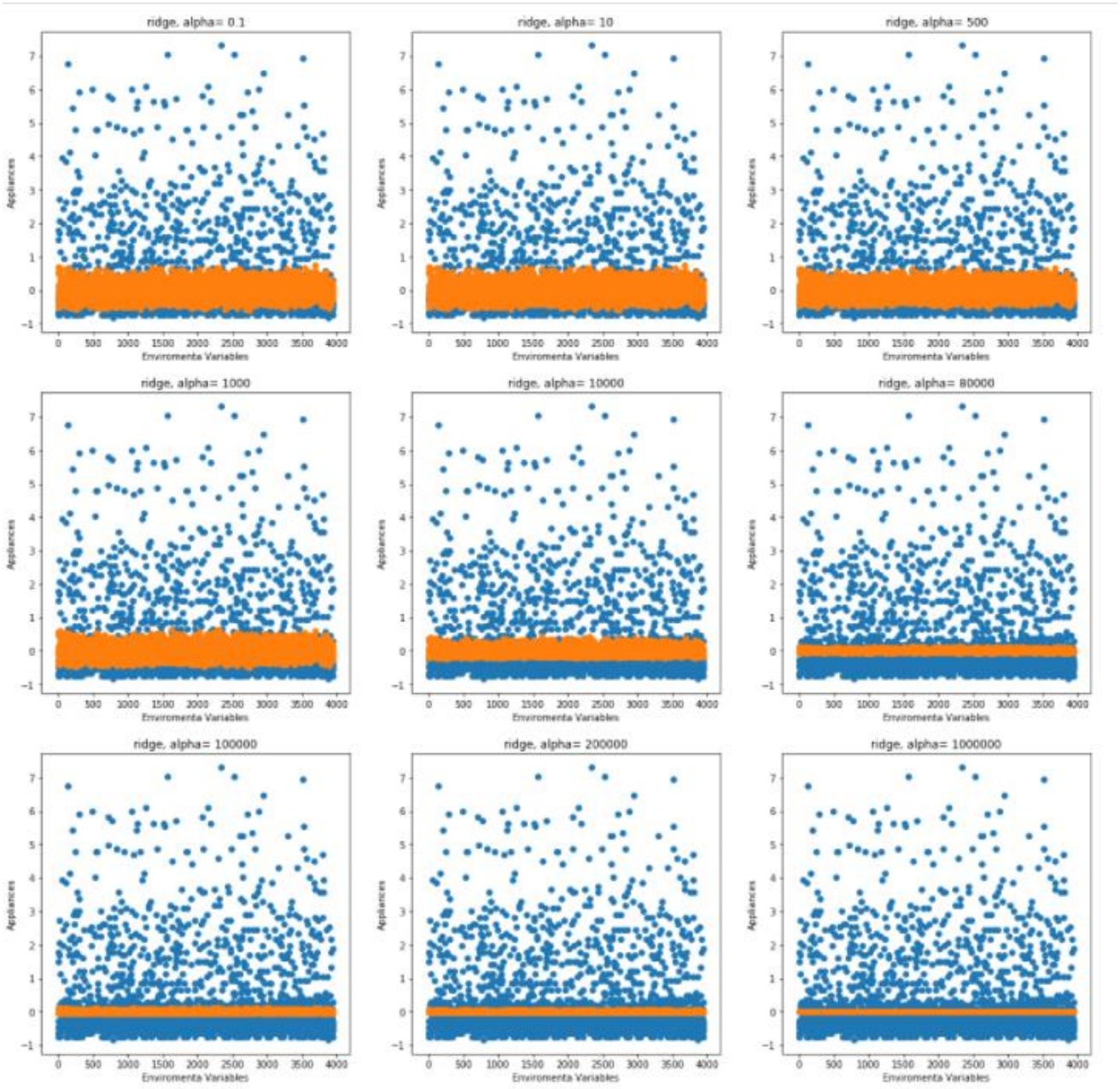

Figure 11 Ridge Regression with different alpha values

Looking at the plots and numerical results, we can notice how the error decrease from the linear regression and it seems like alpha = 0.1 may yield the best results.

7. Quality analysis.

The determination coefficient was shown already in each case individually. In the following figure, we show the distribution of residuals, test the normality distribution of the residuals and plot biplot for each case of regression.

In figure (12), there is the residual distribution in the case of Ridge regression, and the Biplot.

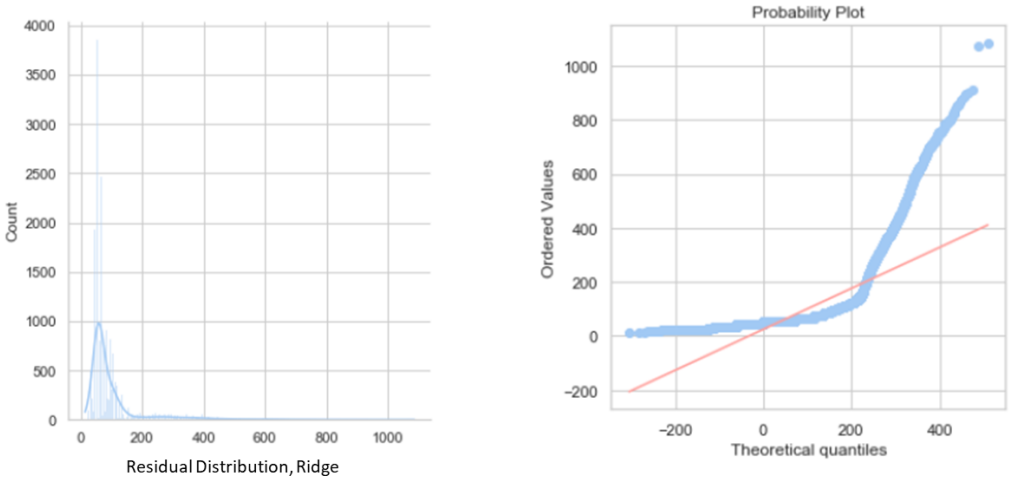

Figure 12 Ridge Residuals

In figure (13), there is the residual distribution in the case of Lasso regression, and the Biplot.

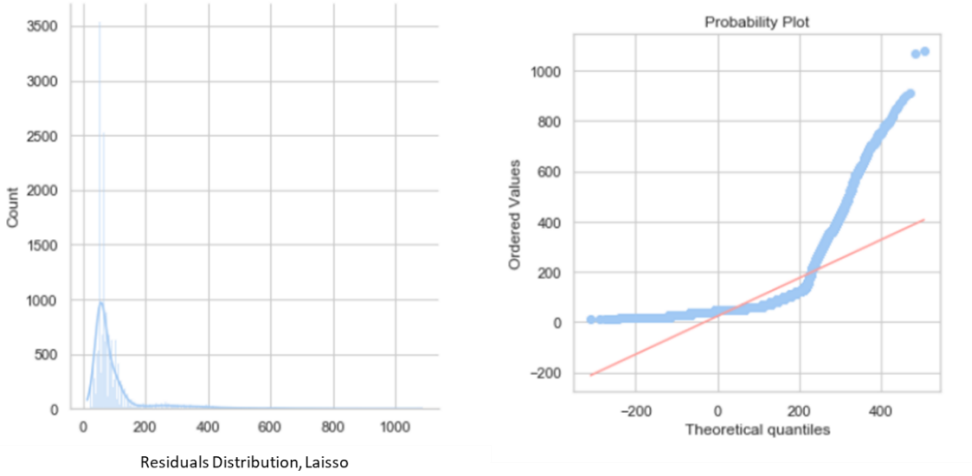

Figure 13 Laisso Residuals

In figure (14), there is the residual distribution in the case of Linear regression, and the Biplot.

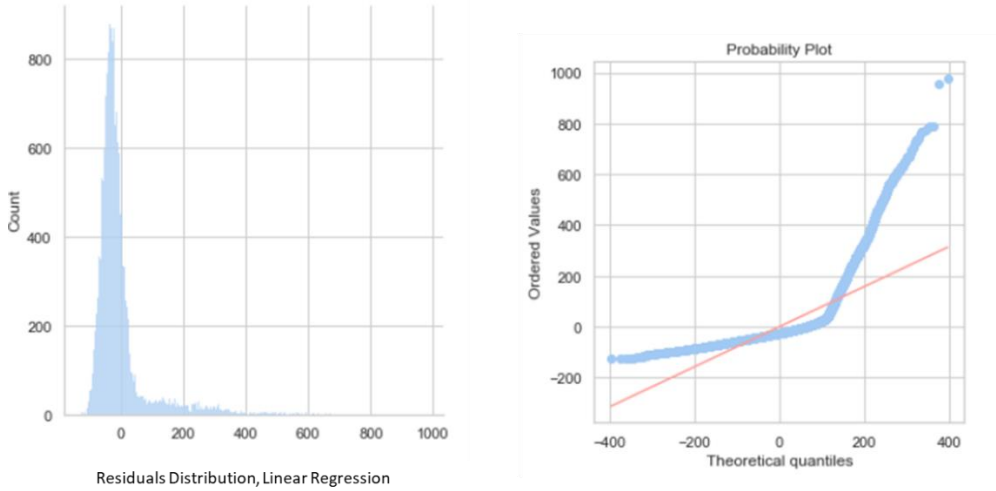

Figure 14 Linear Residuals

In the following table we can find the values of Mean of the residuals, as well as the p-value from Shapiro-Wilk test and Kolmogorov-Smirnov test.

	Mean	Shapiro-Wilk Test	Kolmogorov-Smirnov Test
Linear Regression	0.36150026603017166	0.0	0.0
Laisso	97.6949581960983	0.0	0.0
Ridge	102.64093779621092	0.0	0.0

For the Laisso and Ridge, we can see that the mean value is quite large, which means that the predicted values are larger than the real data. Meanwhile, the mean for the residuals of the Linear Regression is close to zero, which means that the predicted and original data are closer to each other.
We can see that the p-value is small for all the regression types, which means that the normal distribution does not fit the residuals properly, as can be indicated from the biplot of the residuals for each case.
We know that in order to make a valid inference from regression, the residuals of the regression should follow a normal distribution. But with the values of p-values we have and the biplot, we know that none of the residuals are properly following a normal distribution, and since the condition of the normality of the residuals is not satisfied, then the models do not fit the data and the linear regression is not applicable.

8.Source code

Please find my code in the following GitHub link:

https://github.com/neematAllosh/MultiderivativeDataAnalysis/blob/master/lab_02.ipynb