EDA for Particle Identification

By: Adam Erler
Data From Kaggle

EDA Focus

For our EDA we will be exploring only 4 features and see how they relate to a 5th. Our explanatory variables will be

SpdE - energy deposit associated to the track in the Spd

PrsE - energy deposit associated to the track in the Prs

EcalE - energy deposit associated to the track in the Ecal

HcalE - energy deposit associated to the track in the Hcal

Our dependent variable will be the label feature.

Spd stands for Scintillating Pad Detector, PrsE - Preshower, Ecal - electromagnetic calorimeter, Hcal - hadronic calorimeter

Variable Description

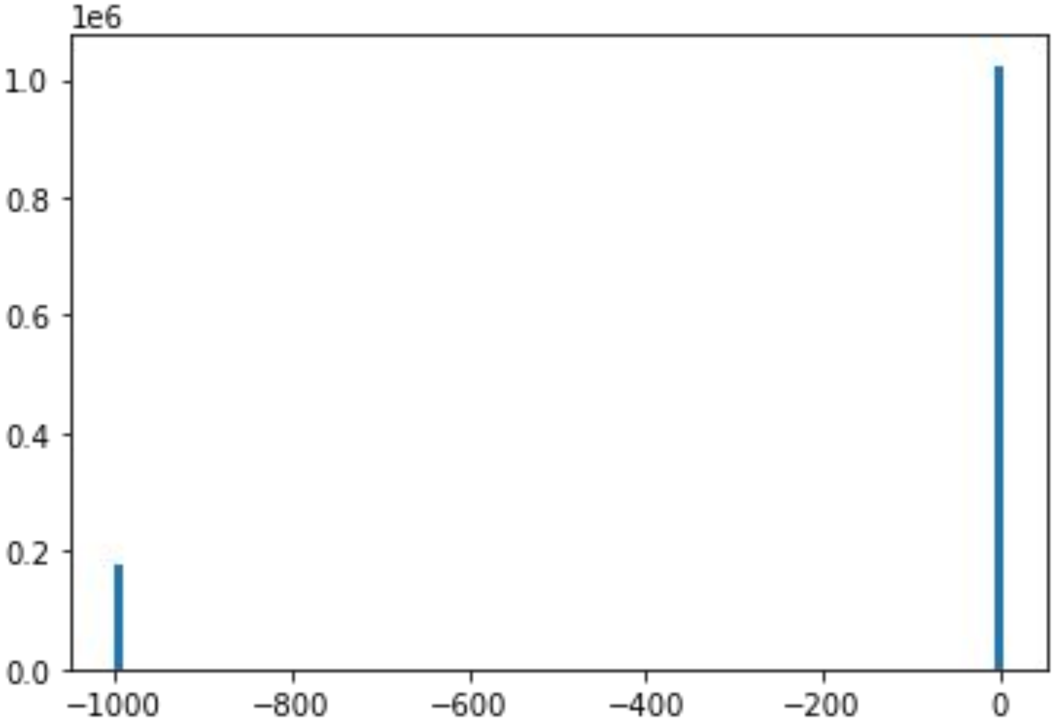
Variable	Description
EcalE	Energy deposited in the Electromagnetic Calorimeter Spd stands for Scintillating Pad Detector, Prs - Preshower, Ecal - electromagnetic calorimeter, Hcal - hadronic calorimeter
HcalE	Energy deposited in the Hadronic calorimeter
PrsE	Preshower energy measured
SpdE	Energy in the scintillating pad detector
Label	Either: Muon, Ghost, Pion, Proton, Kaon, Electron.

Objective

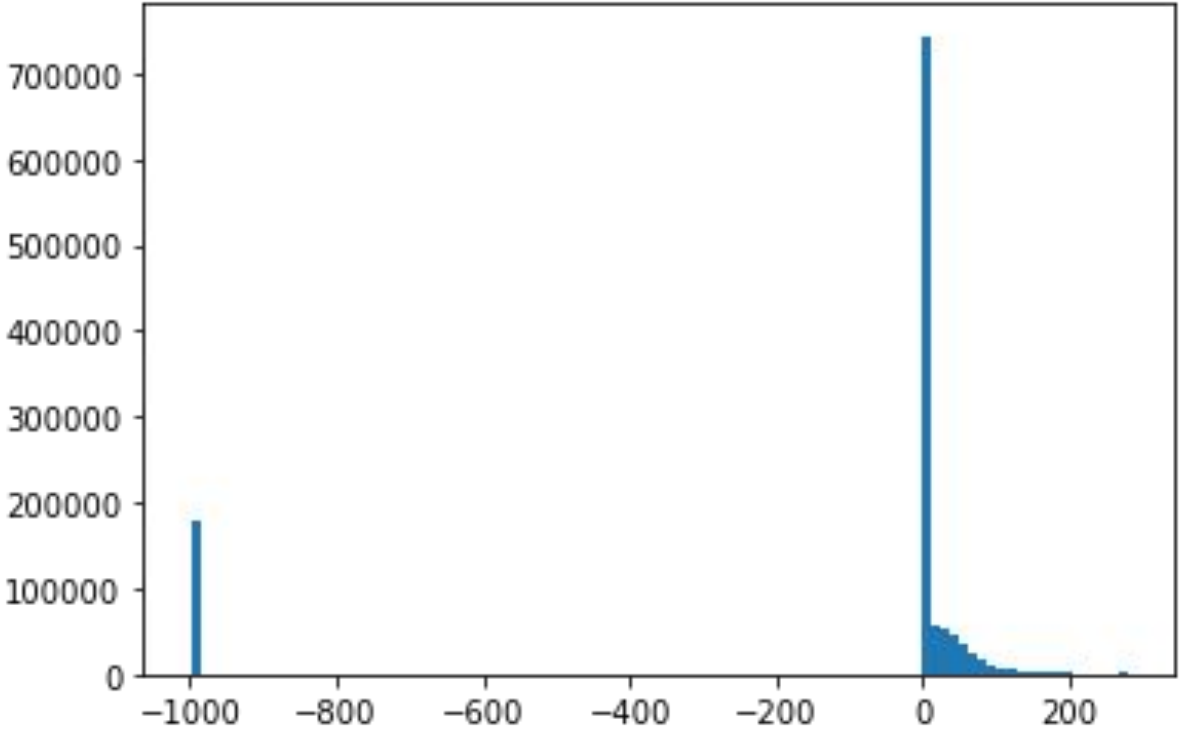
Our goal of this project is to see if we can use the features chosen to explain the label of particles in the detector.

Our null Hypothesis is that the measurements of EcalE, HcalE, PrsE, and SpdE in a detector can be used to correctly identify the type of particle 99% of the time.

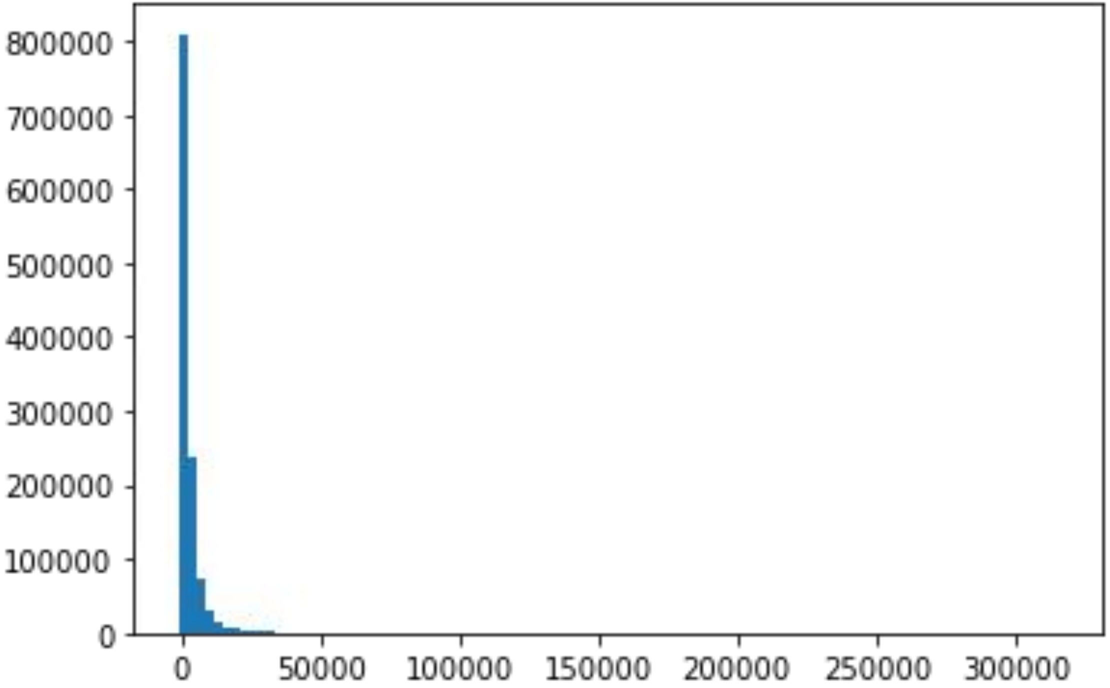
Histograms of Variables -- SpdeE



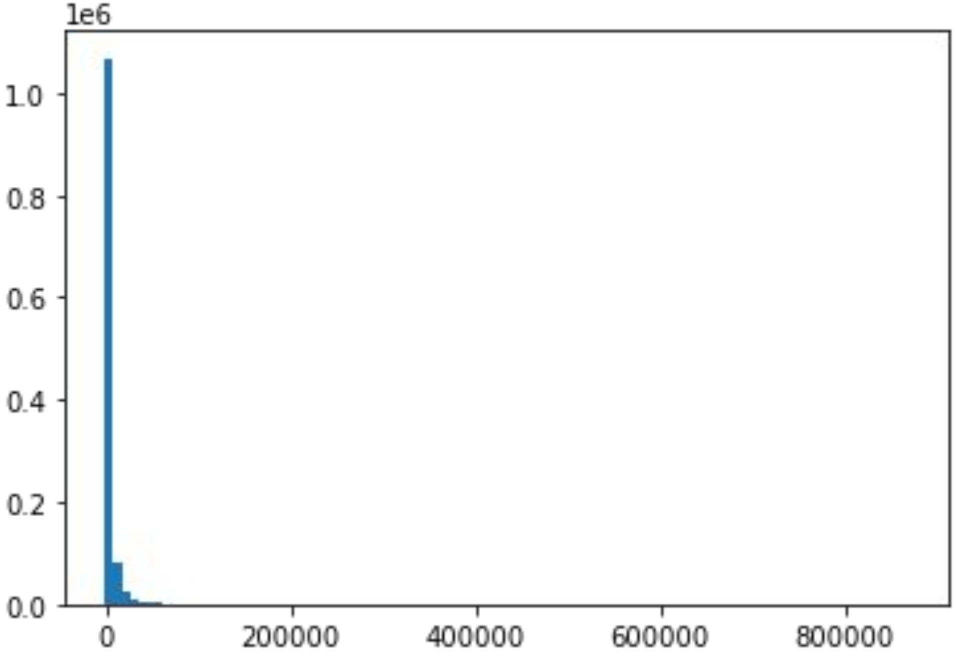
Histograms of Variables -- PrsE



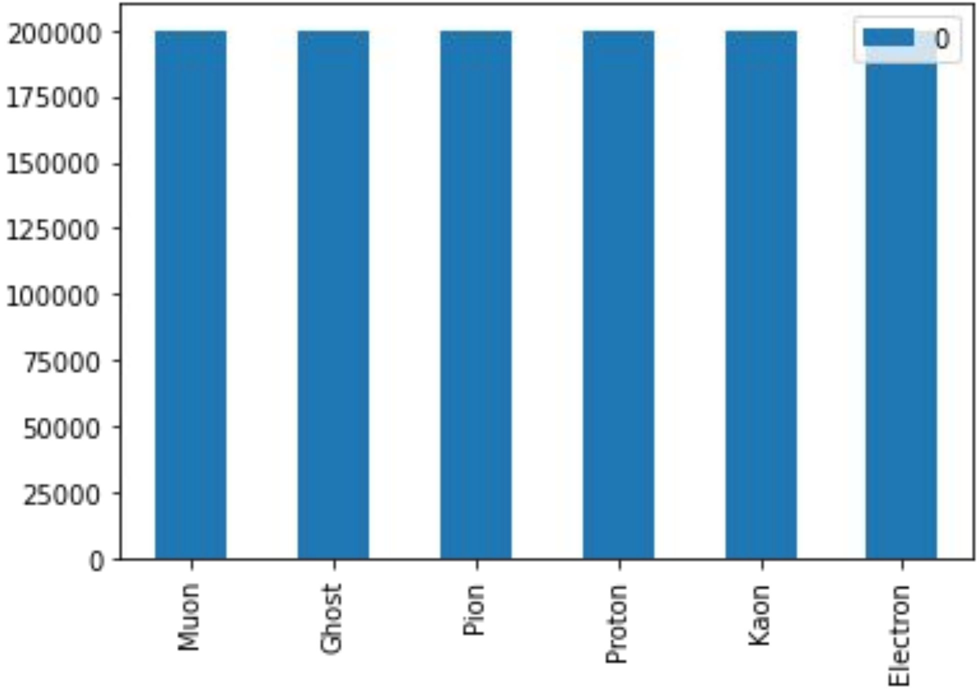
Histograms of Variables -- EcalE



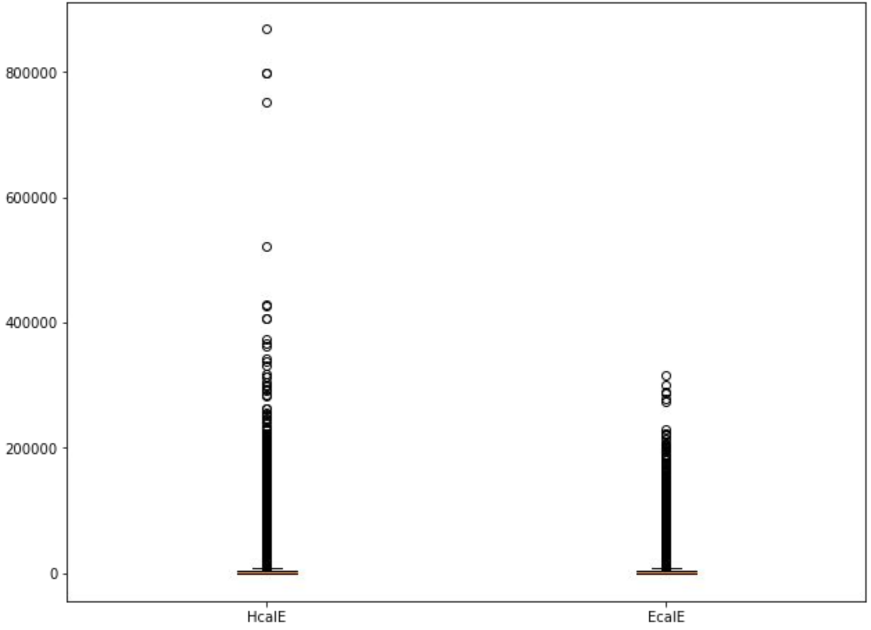
Histograms of Variables -- HcalE

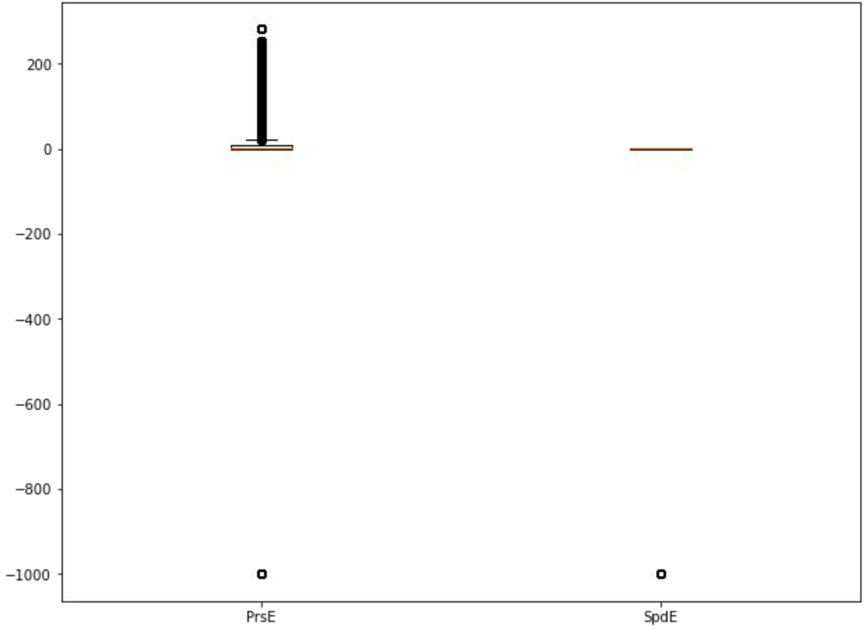


Histograms of Variables -- Label



Box Plot for Explanatory Values





Summary of Statistics

Variable	Mean	Mode	Median	Variance	Skew
SpdE	-144.38	3.20	3.20	125672.75	-2.00
PrsE	-133.89	-999.0	2.47	131694.75	-1.94
EcalE	2346.44	-999.0	659.10	32719796.24	8.45
HcalE	2900.03	-999.0	578.01	59081456.15	11.58

Summary of Descriptive stats.

We can see the values SpdE and PrsE have similar distributions, along with HcalE and EcalE.

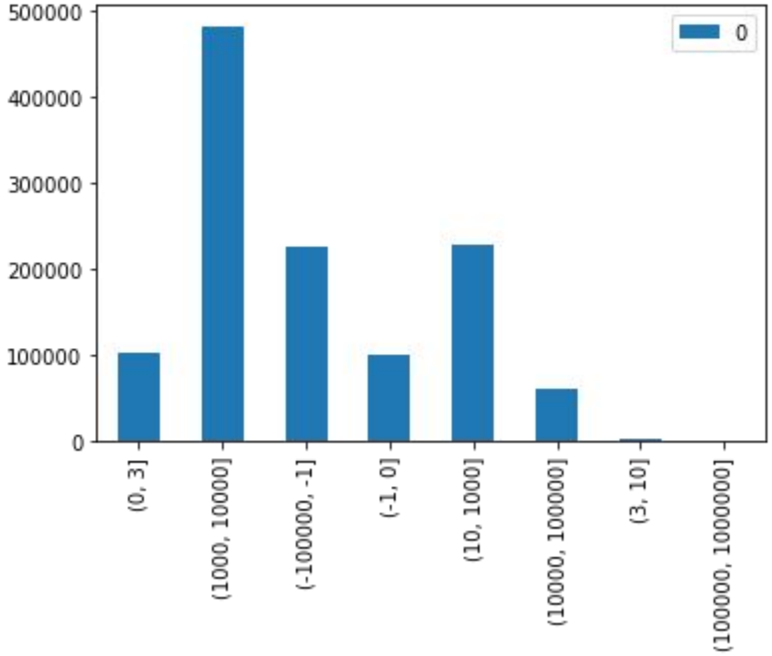
The Label feature appears by design to be uniform in its distribution.

Since the goal of this project is to eventually expand this into a predictive model this distribution is logical.

Looking at the box and whisker plots we can see some clear outliers. In some contexts these would be concerning but since we measuring physical phenomena these values could be useful in identifying new or extreme physics.

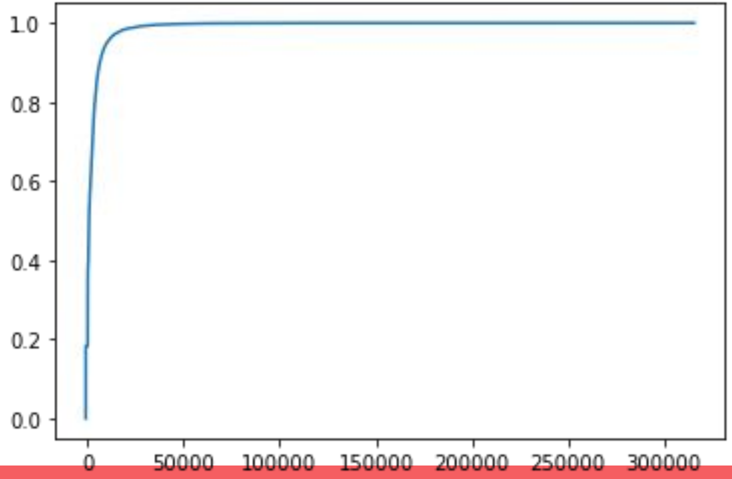
PMF of EcalE

From this we can see most values expect to positive. To fully understand the impact of this we should consider what a negative meaning in the detector means. Until then we cannot fully interpret the result. This is something to consider for future research.



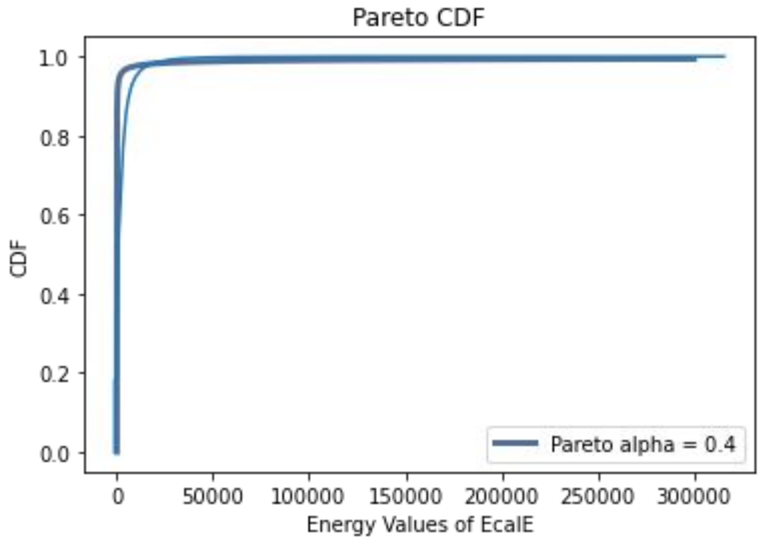
CDF of EcalE

- Create 1 CDF with one of your variables, using page 41-44 as your guide, what does this tell you about your variable and how does it address the question you are trying to answer (Chapter 4).



Analytical Distribution

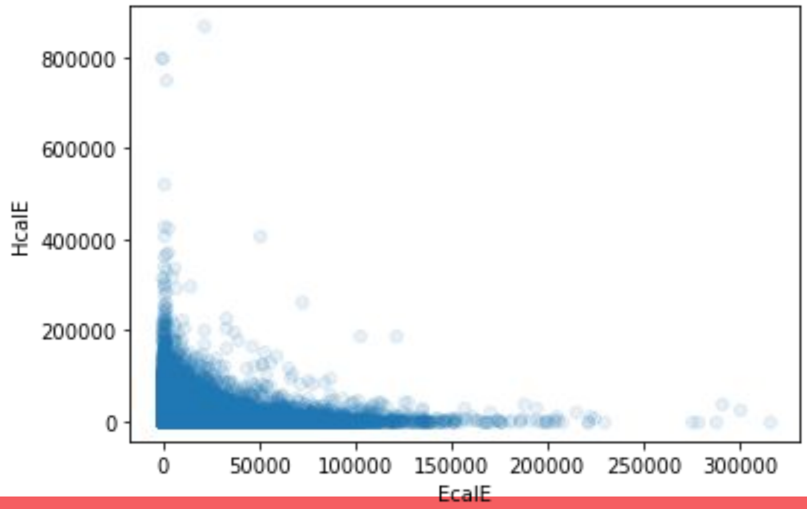
A Pareto Distribution for the energy measurement of particles decaying in the Electromagnetic Calorimeter is one that makes natural sense in the context. This informs me that is likely accurate data of a physical phenomena.



Scatter Plot 1 HcalE vs EcalE

The Pearson Correlation value is low for both a linear and non-linear measurements leading us to believe these two variables are not correlative.

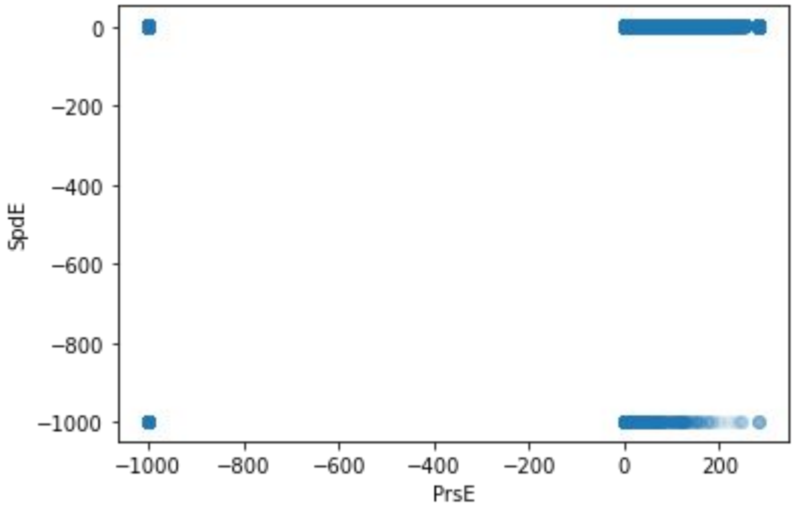
Pearson Correlation	0.0725
Covariance	3187251.941
Non-Linear Pearson ($HcalE^2$)	0.014

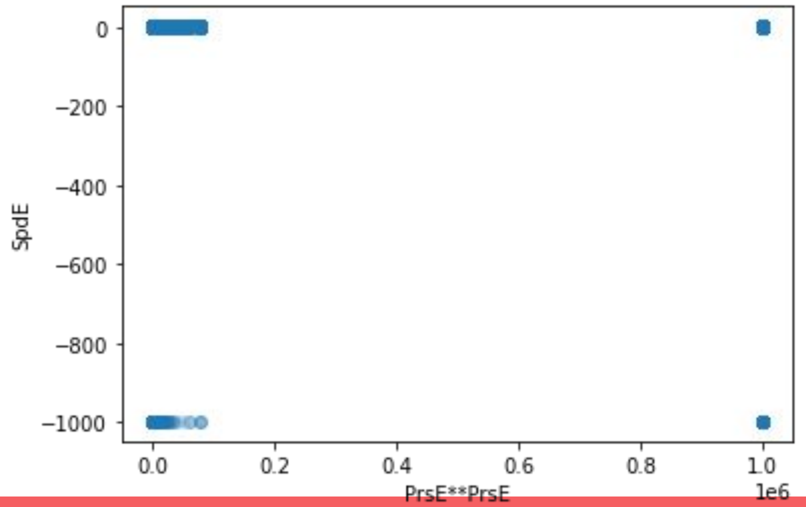


Scatter Plot 2 PrsE vs SpdE

In a linear and non-linear exploration we can see both variables move in a similar pattern. This tells us that if we were doing a feature reduction we could consider dropping one of these.

Pearson Correlation Linear	0.973
Pearson Correlation Non-Linear	-0.977
Covariance Linear	125275
Covariance Non-Linear	-122841408





Hypothesis Test for Variables Explanatory Power

The table below is a list of the permutations of hypothesis tests considering how well each Detectors measurement correlates with weather or not a dummy variable represent a labeled particle. In all cases our p-value was almost 0. We have overwhelming shown our EDA is worthwhile and need to do further study in this relationship.

Detector	Electron	Muon	Ghost	Proton	Kaon	Pion
SpdE	0	0	0	0	0	0
Prse	0	0	0	0	0	0
EcalE	0	0	0	0	0	0
HcalE	0	0	0	0	0	0

Regression of Labels as Dummy Variables

- The next 6 slides will serve as regressions for each of the labels as their own dummy variable.

Ghost Regression

Dep. Variable:	Ghost	**R-squared:**	0.007
Model:	OLS	**Adj. R-squared:**	0.007
Method:	Least Squares	**F-statistic:**	2252.
Date:	Wed, 17 Nov 2021	**Prob (F-statistic):**	0.00
Time:	20:56:12	**Log-Likelihood:**	-5.1379e+05
No. Observations:	1200000	**AIC:**	1.028e+06
Df Residuals:	1199995	**BIC:**	1.028e+06
Df Model:	4		
Covariance Type:	nonrobust		

Omnibus:	335109.090	**Durbin-Watson:**	2.003
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	694598.148
Skew:	1.769	**Prob(JB):**	0.00
Kurtosis:	4.175	**Cond. No.**	1.08e+04

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.1741	0.000	399.610	0.000	0.173	0.175
HcalE	-1.881e-06	4.51e-08	-41.739	0.000	-1.97e-06	-1.79e-06
EcalE	-3.345e-06	6.22e-08	-53.733	0.000	-3.47e-06	-3.22e-06
PrsE	7.133e-06	4.18e-06	1.705	0.088	-1.07e-06	1.53e-05
SpdE	-4.756e-05	4.25e-06	-11.200	0.000	-5.59e-05	-3.92e-05

Proton Regression

Dep. Variable:	Proton	**R-squared:**	0.007
Model:	OLS	**Adj. R-squared:**	0.007
Method:	Least Squares	**F-statistic:**	2182.
Date:	Wed, 17 Nov 2021	**Prob (F-statistic):**	0.00
Time:	20:57:53	**Log-Likelihood:**	-5.1393e+05
No. Observations:	1200000	**AIC:**	1.028e+06
Df Residuals:	1199995	**BIC:**	1.028e+06
Df Model:	4		
Covariance Type:	nonrobust		

Omnibus:	336115.134	**Durbin-Watson:**	2.000
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	698240.988
Skew:	1.769	**Prob(JB):**	0.00
Kurtosis:	4.204	**Cond. No.**	1.08e+04

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.1629	0.000	373.894	0.000	0.162	0.164
HcalE	3.452e-06	4.51e-08	76.596	0.000	3.36e-06	3.54e-06
EcalE	-1.405e-06	6.23e-08	-22.572	0.000	-1.53e-06	-1.28e-06
PrsE	-0.0002	4.18e-06	-39.635	0.000	-0.000	-0.000
SpdE	0.0002	4.25e-06	40.968	0.000	0.000	0.000

Pion Regression

Dep. Variable:	Pion	**R-squared:**	0.006
Model:	OLS	**Adj. R-squared:**	0.006
Method:	Least Squares	**F-statistic:**	1797.
Date:	Wed, 17 Nov 2021	**Prob (F-statistic):**	0.00
Time:	21:18:23	**Log-Likelihood:**	-5.1469e+05
No. Observations:	1200000	**AIC:**	1.029e+06
Df Residuals:	1199995	**BIC:**	1.029e+06
Df Model:	4		
Covariance Type:	nonrobust		

Omnibus:	335670.612	**Durbin-Watson:**	1.997
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	696539.454
Skew:	1.771	**Prob(JB):**	0.00
Kurtosis:	4.179	**Cond. No.**	1.08e+04

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.1704	0.000	391.008	0.000	0.170	0.171
HcalE	2.18e-06	4.51e-08	48.351	0.000	2.09e-06	2.27e-06
EcalE	-3.165e-06	6.23e-08	-50.800	0.000	-3.29e-06	-3.04e-06
PrsE	-0.0002	4.19e-06	-37.672	0.000	-0.000	-0.000
SpdE	0.0002	4.25e-06	38.762	0.000	0.000	0.000

Electron Regression

Omnibus:	335670.612	**Durbin-Watson:**	1.997
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	696539.454
Skew:	1.771	**Prob(JB):**	0.00
Kurtosis:	4.179	**Cond. No.**	1.08e+04

Dep. Variable:	Pion	**R-squared:**	0.006
Model:	OLS	**Adj. R-squared:**	0.006
Method:	Least Squares	**F-statistic:**	1797.
Date:	Wed, 17 Nov 2021	**Prob (F-statistic):**	0.00
Time:	21:18:23	**Log-Likelihood:**	-5.1469e+05
No. Observations:	1200000	**AIC:**	1.029e+06
Df Residuals:	1199995	**BIC:**	1.029e+06
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.1704	0.000	391.008	0.000	0.170	0.171
HcalE	2.18e-06	4.51e-08	48.351	0.000	2.09e-06	2.27e-06
EcalE	-3.165e-06	6.23e-08	-50.800	0.000	-3.29e-06	-3.04e-06
PrsE	-0.0002	4.19e-06	-37.672	0.000	-0.000	-0.000
SpdE	0.0002	4.25e-06	38.762	0.000	0.000	0.000

Muon Regression

Omnibus:	319858.880	**Durbin-Watson:**	2.001
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	643288.609
Skew:	1.711	**Prob(JB):**	0.00
Kurtosis:	4.077	**Cond. No.**	1.08e+04

Dep. Variable:	Muon	**R-squared:**	0.033
Model:	OLS	**Adj. R-squared:**	0.033
Method:	Least Squares	**F-statistic:**	1.038e+04
Date:	Wed, 17 Nov 2021	**Prob (F-statistic):**	0.00
Time:	21:18:24	**Log-Likelihood:**	-4.9786e+05
No. Observations:	1200000	**AIC:**	9.957e+05
Df Residuals:	1199995	**BIC:**	9.958e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.2109	0.000	490.575	0.000	0.210	0.212
HcalE	-1.536e-06	4.45e-08	-34.551	0.000	-1.62e-06	-1.45e-06
EcalE	-1.121e-05	6.14e-08	-182.570	0.000	-1.13e-05	-1.11e-05
PrsE	-0.0002	4.13e-06	-38.631	0.000	-0.000	-0.000
SpdE	0.0002	4.19e-06	57.474	0.000	0.000	0.000

Kaon Regression

Omnibus:	334886.038	**Durbin-Watson:**	1.999
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	694384.352
Skew:	1.758	**Prob(JB):**	0.00
Kurtosis:	4.234	**Cond. No.**	1.08e+04

Dep. Variable:	Kaon	**R-squared:**	0.011
Model:	OLS	**Adj. R-squared:**	0.011
Method:	Least Squares	**F-statistic:**	3272.
Date:	Sat, 20 Nov 2021	**Prob (F-statistic):**	0.00
Time:	10:53:06	**Log-Likelihood:**	-5.1177e+05
No. Observations:	1200000	**AIC:**	1.024e+06
Df Residuals:	1199995	**BIC:**	1.024e+06
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.1615	0.000	371.327	0.000	0.161	0.162
HcalE	4.376e-06	4.5e-08	97.296	0.000	4.29e-06	4.46e-06
EcalE	-1.352e-06	6.21e-08	-21.751	0.000	-1.47e-06	-1.23e-06
PrsE	-0.0002	4.18e-06	-40.039	0.000	-0.000	-0.000
SpdE	0.0002	4.24e-06	43.630	0.000	0.000	0.000