ASSIGNMENT 7
UNSUPERVISED LEARNING

PART 1

5420 Anomaly Detection, Fall 2020

- Harsh Dhanuka, hd2457

Agenda

HEALTHCARE FRAUD

Hospitals **Insurance**

Detect specific transaction which are suspicious

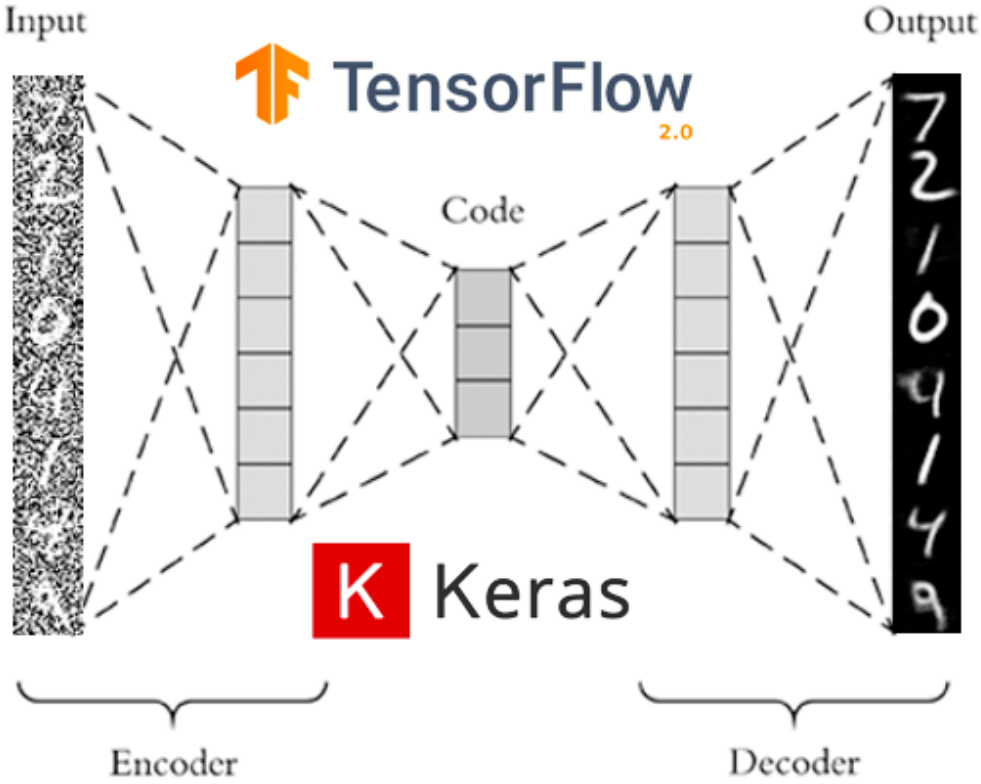
Techniques? Clustering? What kind of clustering?

How to judge fraud? How to justify results?

Autoencoder Clustering

An autoencoder is a special type of neural network that copies the input values to the output values. It does not require the target variable like the conventional Y, thus it is categorized as unsupervised learning.

If the number of neurons in the hidden layers is less than that of the input layers, the hidden layers will extract the essential information of the input values. This condition forces the hidden layers to learn the most patterns of the data and ignore the "noises".



So in an autoencoder model, the hidden layers must have fewer dimensions than those of the input or output layers. If the number of neurons in the hidden layers is more than those of the input layers, the neural network will be given too much capacity to learn the data.

Considerations

1. Removed features

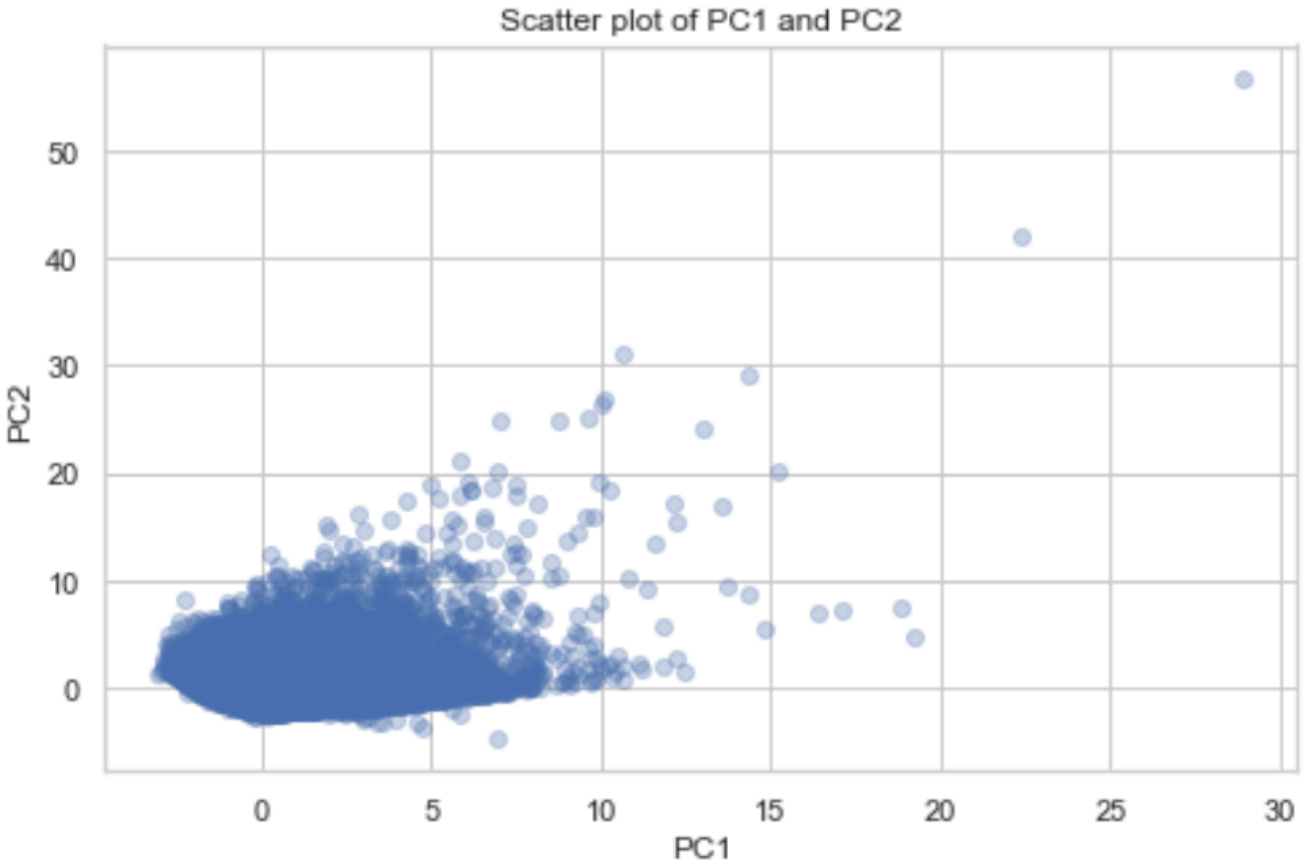
- Removed benchmark features such as ‘State Total’ and ‘Total Discharge’ from the scaled dataset

2. Initial PCA plot to check data

- Plot a 2 component initial PCA plot, to check data distribution

3. Split to train_test:

- 75% split, train has 75% of the data.
- Now, for the test data, I will be using the **entire 100% data**, as even the train data has anomalies.



Autoencoder Clustering

1 Model

Build 3 models, with different levels of hidden layers. Check model stability using the 'average' aggregate method:

[6, 5, 5, 6]

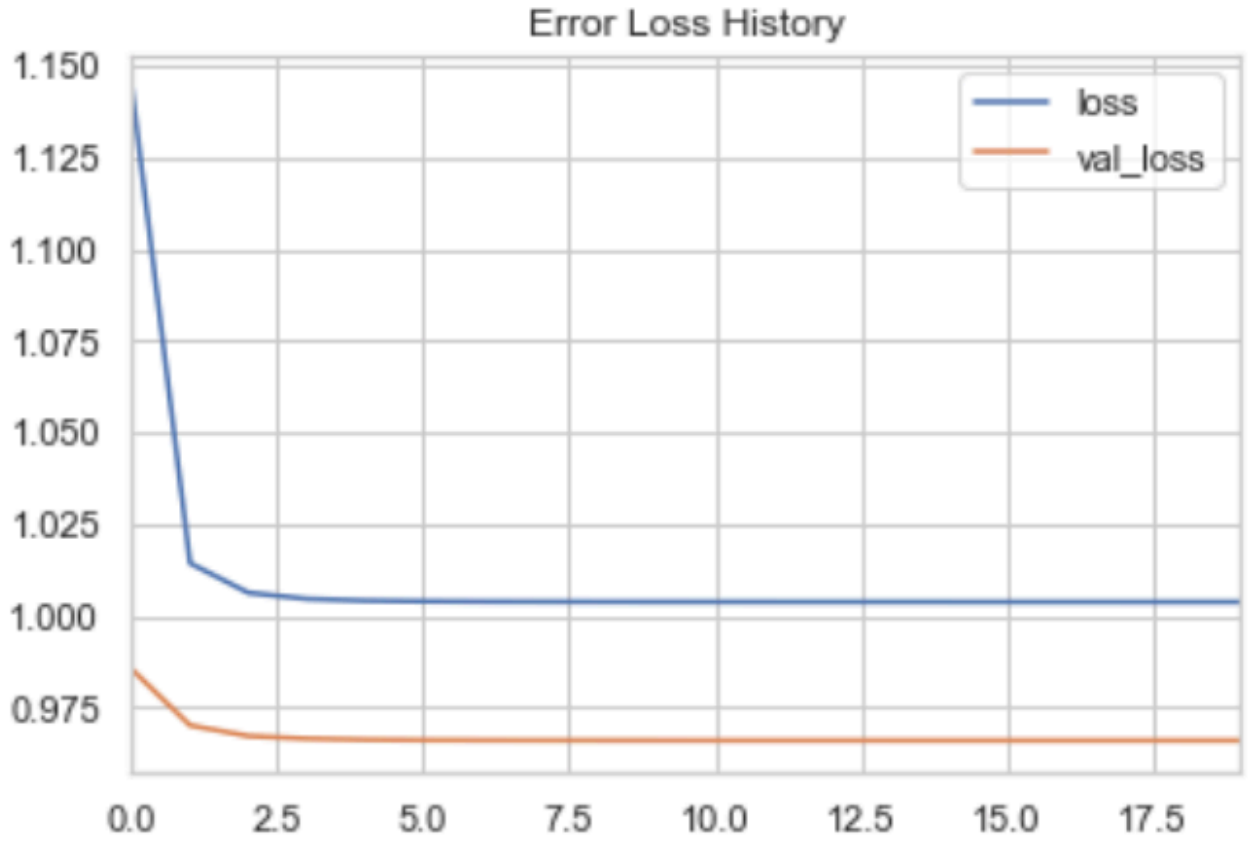
[6, 5, 2, 5, 6]

[6, 5, 3, 2, 3, 5, 6]

2 Visualize the Loss

The goal of model training is to minimize the loss. This loss describes the objective that the autoencoder tries to reach.

Epochs = 20 Contamination = 0.1 or 10%

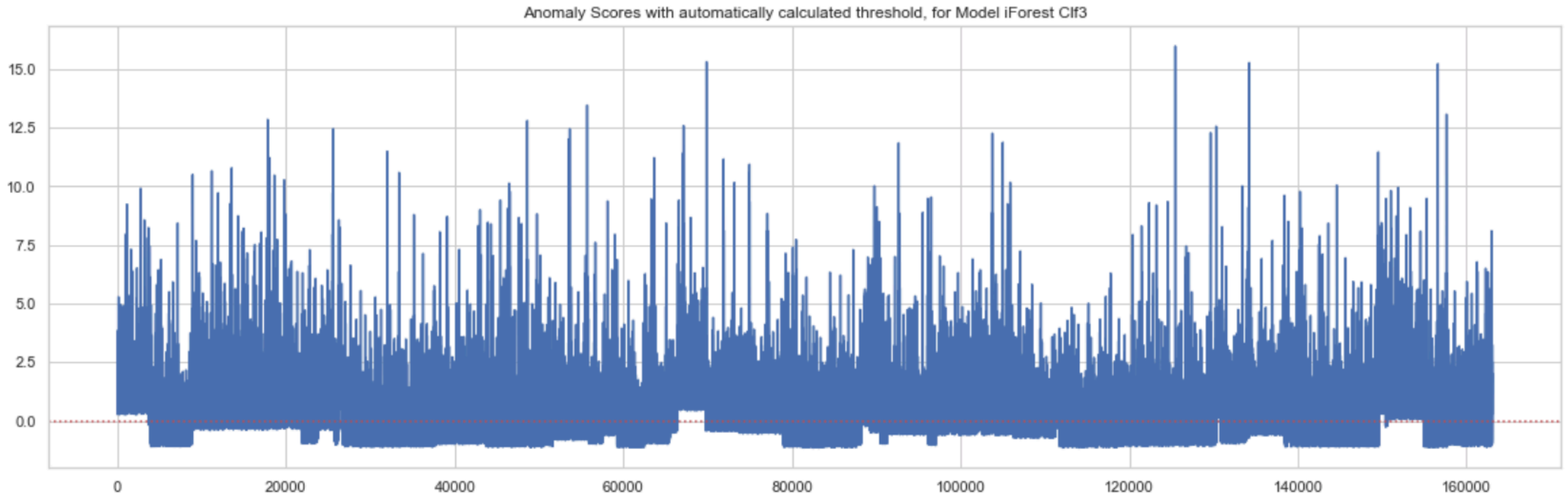


One Epoch is when an ENTIRE dataset is passed forward and backward through the neural network only ONCE

Autoencoder Clustering

3 Plotting all scores

Anomaly distances, with automatically calculated threshold, as per the Model 3 algorithm



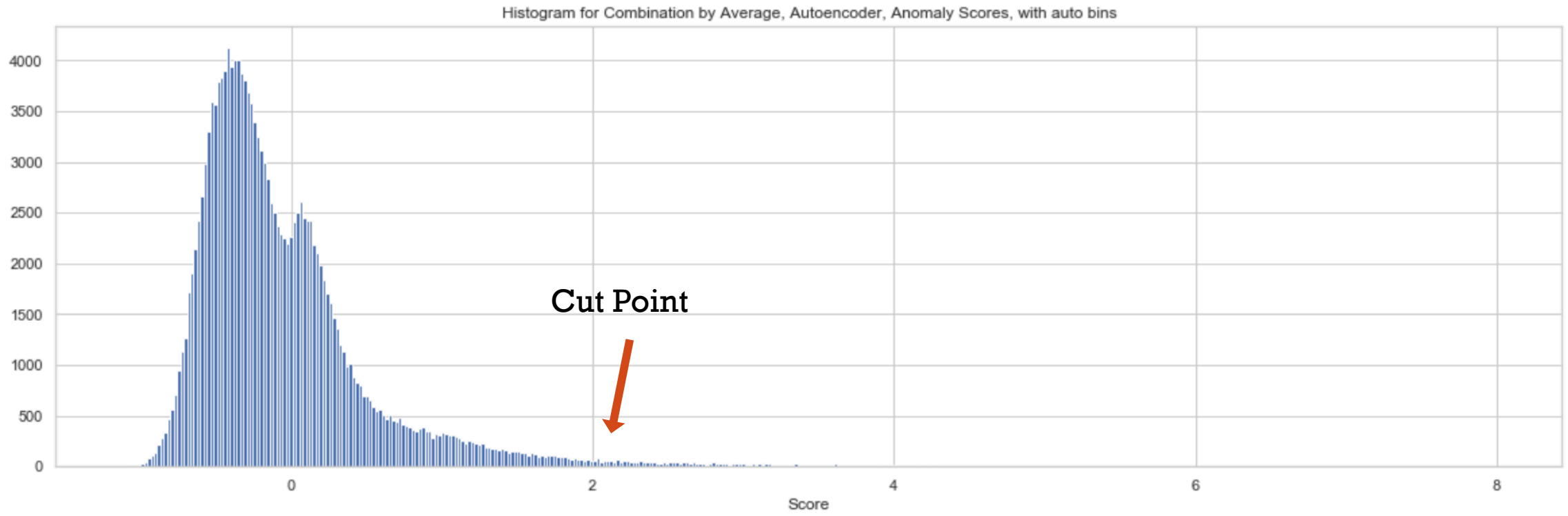
Autoencoder Clustering

4 **Reasonable Boundaries**

I will chose 2 different cut points, which are:

2.0

10.0 This will result in a 3 cluster analysis



Autoencoder Clustering

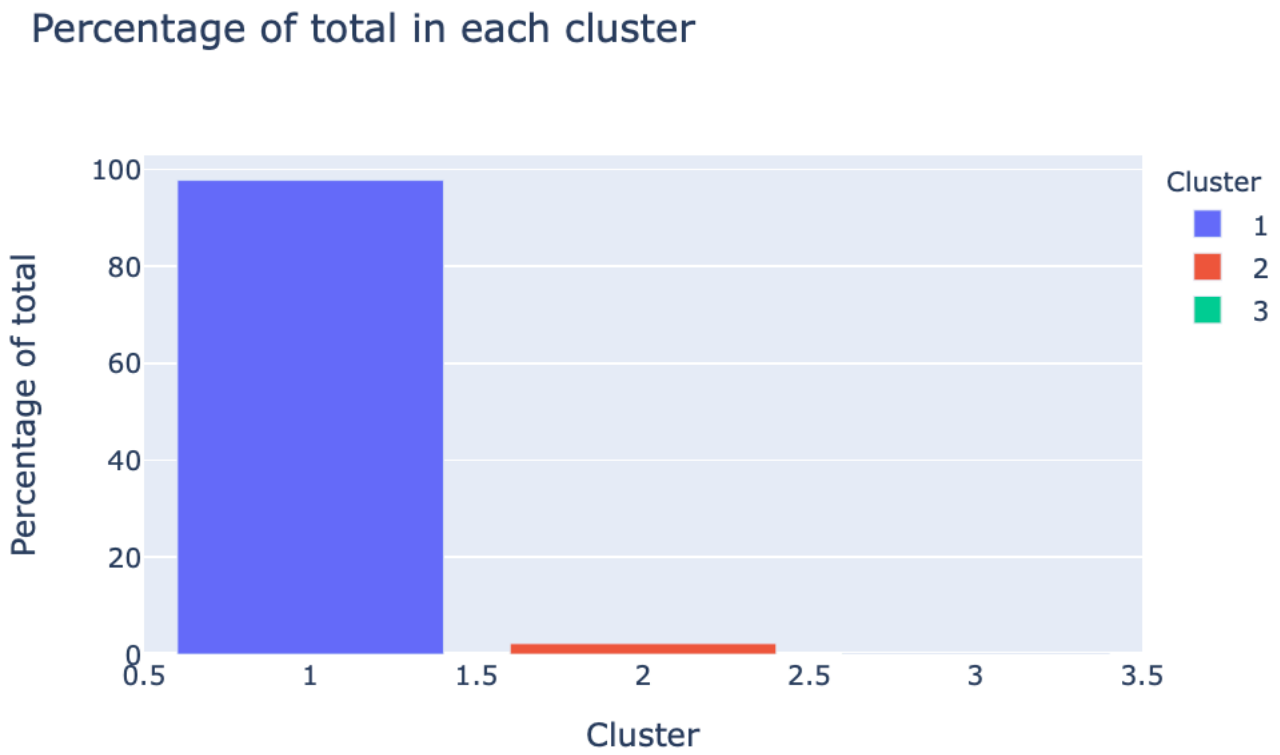
5 **Clusters**

Check the statistics of the 3 clusters.

Here, I am showing the percentage of data points in each cluster

Cluster	Data points
1	159530
2	3405
3	130

Percentage of total	Cluster
97.832153	1
2.088124	2
0.079723	3



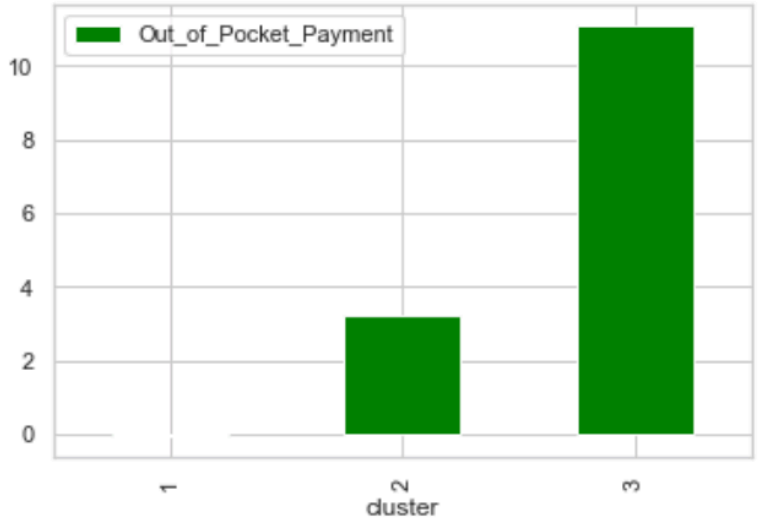
Cluster Evaluation

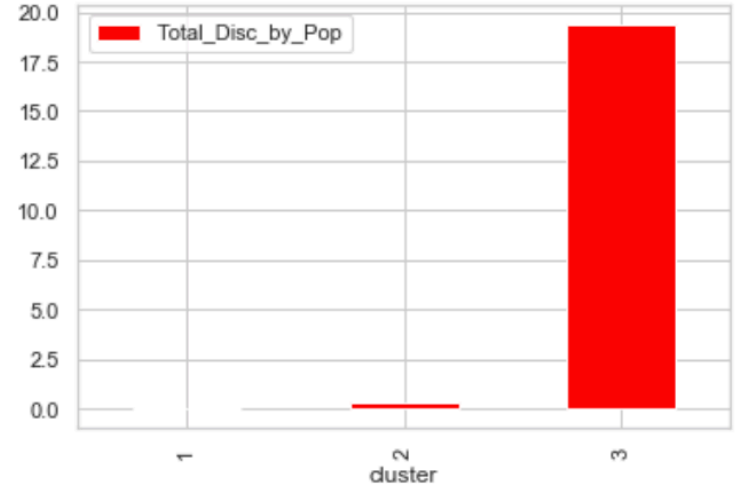
Cluster wise variable Averages

6 **Cluster 2 and 3 have less than 5% of the total data points**

Out of these two clusters, cluster 3 has extremes or high standard deviation from mean for some variables, and hence, I will consider this cluster as suspicious.

Feature-wise cluster mean EDA will be imperative to justify this claim. On the right, I show the same for two features: 'Out of Pocket Payment' and 'Total Discharges by Zipcode Population'





cluster	Average_Total_Payments	Medicare_%_Paid	Medicare_%_Paid_State	Out_of_Pocket_Payment	Median_Score	Total_Disc_by_Pop
1	-0.066432	0.022539	-0.004131	-0.080353	-0.043881	-0.023706
2	2.984325	-0.980106	0.171614	3.344481	1.907198	0.364024
3	3.355793	-1.987412	0.574058	11.005405	3.894464	19.556036

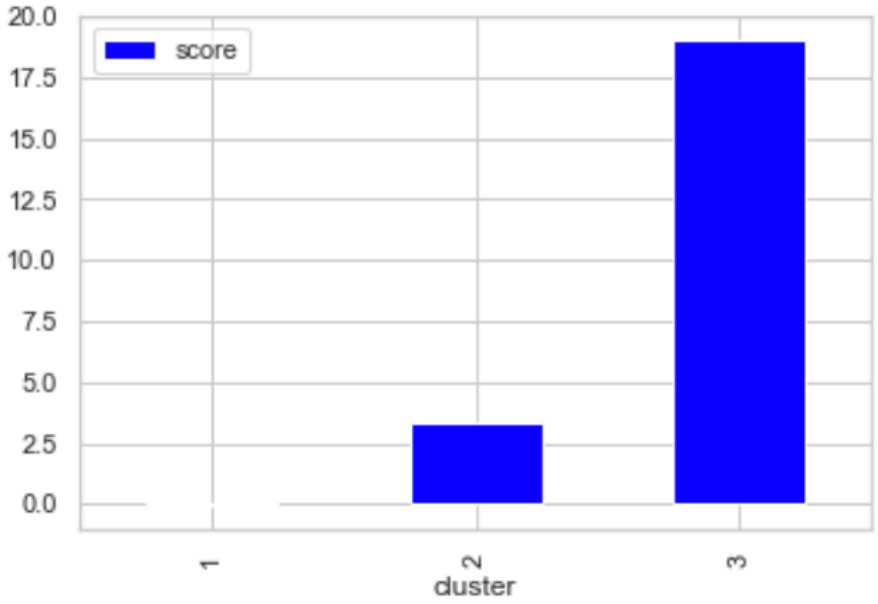
Business Insight

7 Anomaly Score

Anomaly Score gives us insights about the clusters which are potential anomalies, as the those clusters will have a very high anomaly score compared to others.

Cluster 3 has a score almost 19 times higher than cluster 1 and 6 times higher than cluster 2. So, I can safely conclude that Cluster 3 is a potential anomaly.

So, I would pass on the 130 specific entries of the Cluster 3 to the relevant authorities, and call for further investigation on each of the entries, to understand of they are true anomalies. I will provide all the reasoning as I have highlighted above, as to the differences in the means, and walk through the process I have done.



cluster	score
1	-0.090435
2	3.363332
3	19.391270

iForest Clustering

Isolation Forest ot iForest is an anomaly detection algorithm created by Fei Tony Liu et al. They argue that most of the existing approaches to anomaly detection find the norm first, then identify observations that do not conform to the norm. They propose the Isolation Forest as an alternative approach — explicitly isolating anomalies instead of profiling normal data points.

Anomalies are isolated closer to the root of the tree; whereas normal points are isolated at the deeper end of the tree. They call each tree the Isolation Tree or iTree. This isolation characteristic of tree forms the basis to detect anomalies.

It is important to note that this iTree algorithm is different from the decision tree algorithm because iTree does not use a target variable to train the tree. It is an unsupervised learning method.

iForest Clustering

1 Model

Build 3 models, with different number of *max samples.* Check model stability using the 'average' aggregate method:

max_samples = 100% of length of train
max_samples = 80% of length of train
max_samples = 60% of length of train

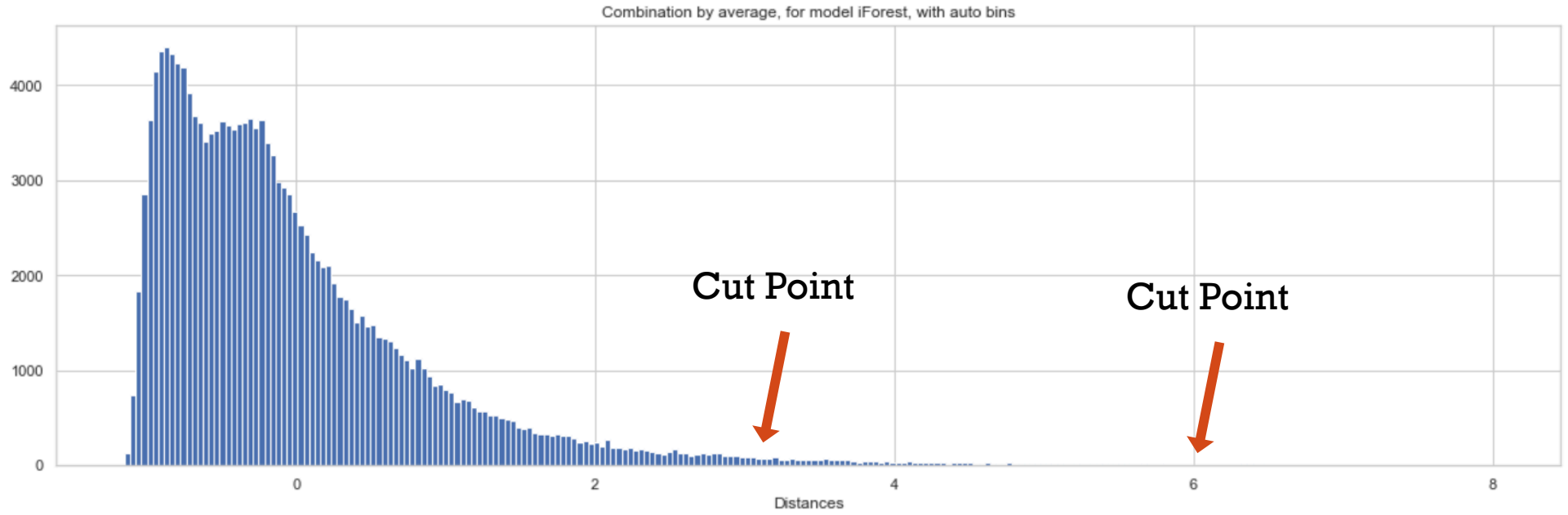
2 Reasonable Boundaries

I will chose 2 different cut points, which are:

3.0

6.0

This will result in a 3 cluster analysis



iForest Clustering

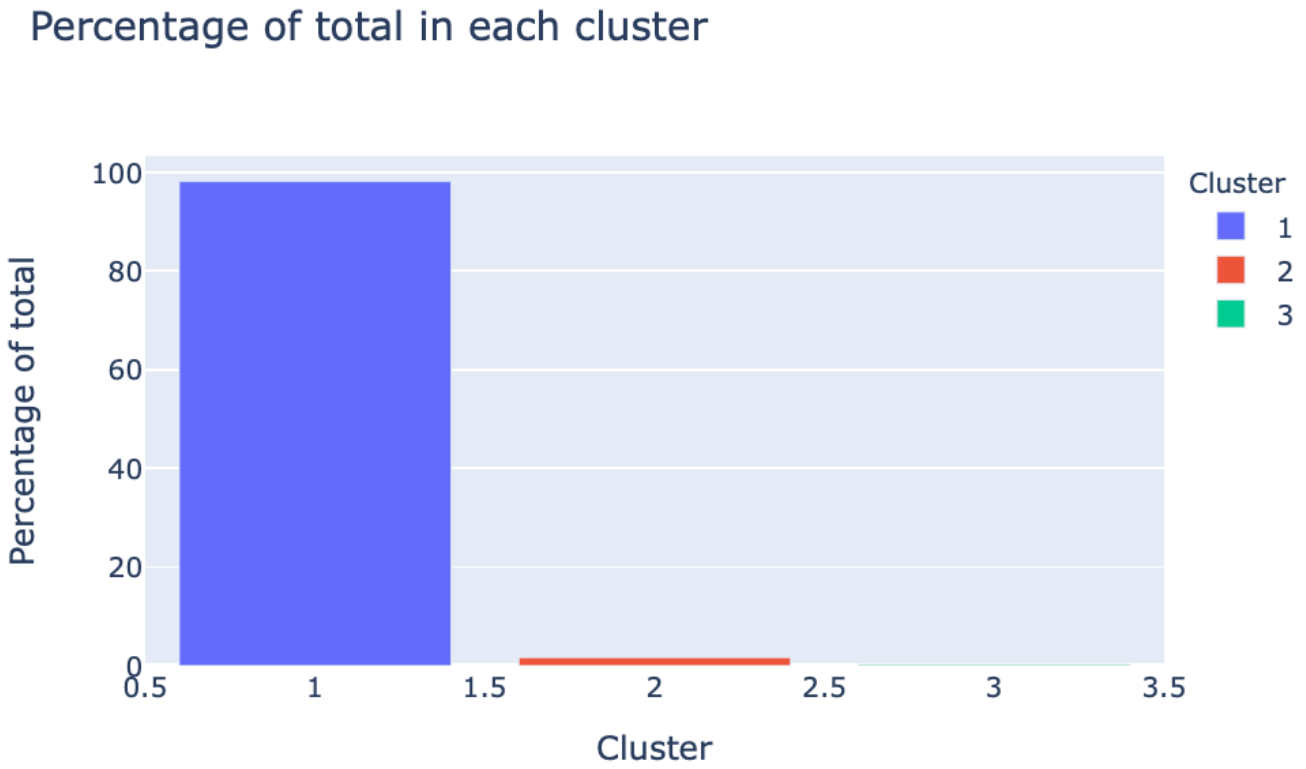
3 **Clusters**

Check the statistics of the 3 clusters.

Here, I am showing the percentage of data points in each cluster

Cluster	Data points
1	159987
2	2683
3	395

Percentage of total	Cluster
98.112409	1
1.645356	2
0.242235	3



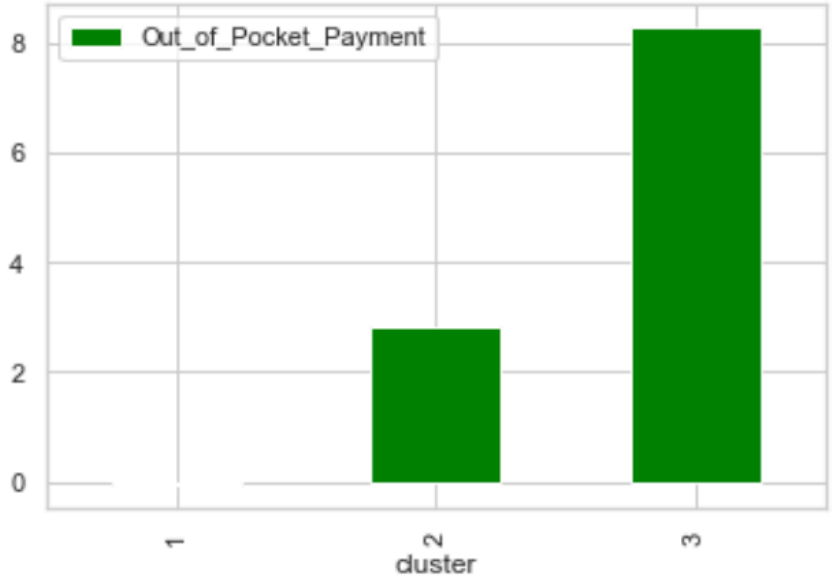
Cluster Evaluation

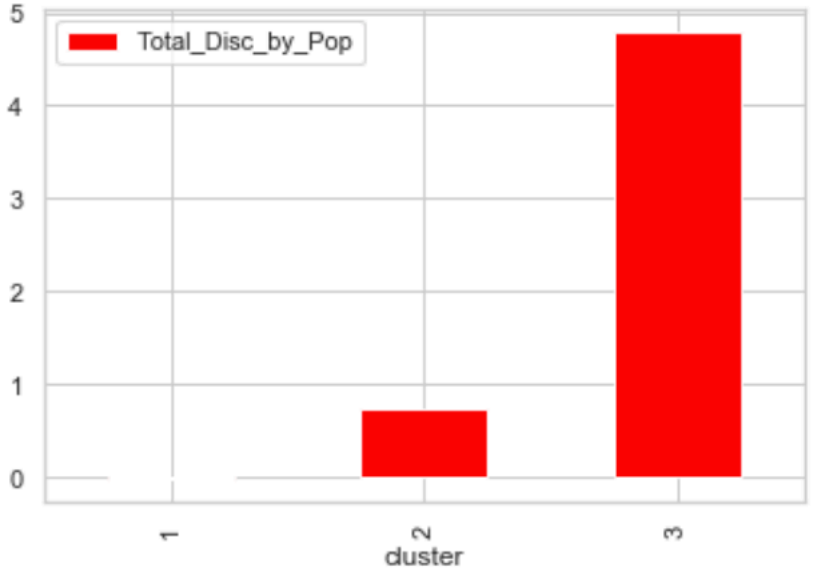
Cluster wise variable Averages

4 **Cluster 2 and 3 have less than 5% of the total data points**

Out of these two clusters, cluster 3 has extremes or high standard deviation from mean for some variables, and hence, I will consider this cluster as suspicious.

Feature-wise cluster mean EDA will be imperative to justify this claim. On the right, I show the same for two features: 'Out of Pocket Payment' and 'Total Discharges by Zipcode Population'





cluster	Average_Total_Payments	Medicare_%_Paid	Medicare_%_Paid_State	Out_of_Pocket_Payment	Median_Score	Total_Disc_by_Pop
1	-0.051218	0.020552	-0.002215	-0.067643	-0.036722	-0.024065
2	2.438619	-0.915684	0.107177	2.811623	1.553160	0.727990
3	4.180836	-2.104292	0.168957	8.299708	4.323665	4.802426

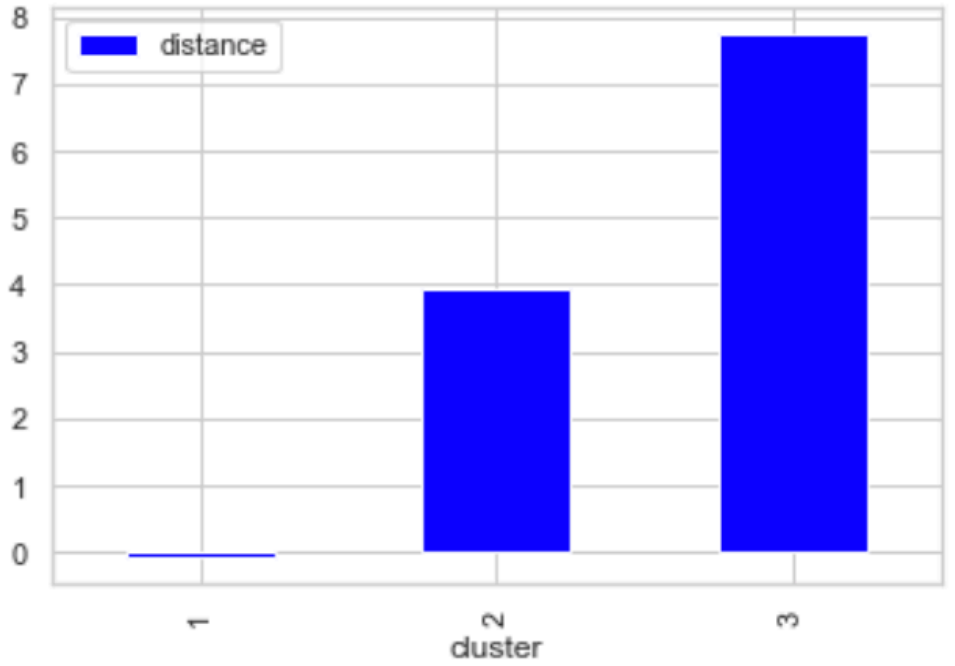
Business Insight

5 Anomaly Distance Score

Anomaly Score gives us insights about the clusters which are potential anomalies, as the those clusters will have a very high anomaly score compared to others.

Cluster 3 has a score almost 7 times higher than cluster 1 and 2.5 times higher than cluster 2. So, I can safely conclude that Cluster 3 is a potential anomaly.

So, I would pass on the 395 specific entries of the Cluster 3 to the relevant authorities, and call for further investigation on each of the entries, to understand of they are true anomalies. I will provide all the reasoning as I have highlighted above, as to the differences in the means, and walk through the process I have done.



cluster	distance
1	-0.083738
2	3.949483
3	7.763719

PyOD Models Comparison

1 kNN

y_by_average_cluster	y_by_average_score
1	-0.211456
2	0.282751
3	1.697698
4	13.632964

256 data points

2 PCA

y_by_average_cluster	y_by_average_score
1	-0.150710
2	1.368969
3	2.935768
4	9.372393

638 data points

3 Autoencoder

cluster	score
1	-0.090435
2	3.363332
3	19.391270

130 data points

4 iForest

cluster	score
1	-0.095299
2	3.375230
3	19.081473

395 data points