Temporal Clustering Of Aggregated PHQ-9 Scores On Sample Dataset

Problem Statement :

The aim of this analysis is to detect significant changes in the Aggregated Patient Health Questionnaire-9 (PHQ-9) scores across 365 days using some aggregated statistics. The sample dataset comprises 100 patients identified by unique patient IDs, with each patient's PHQ-9 score recorded for 50 randomly selected days amongst the total 365 days of the year. Also, each and every patients' are not attempting this survey everyday, rather at most they attempt the survey 6 times during a 365 day span. So, scores of them are sparsed as well.

Dataset Description :

The dataset consists of 100 rows each representing a unique patient ID, and 50 columns representing discrete days. Each cell in the dataset contains the PHQ-9 score of a specific patient on a particular day. The PHQ-9 scores are recorded on a continuous scale ranging from 0 to 27, reflecting the patient's depression severity of depressive symptoms. It is important to note that not all patients respond to the surveys every day, resulting in missing data, represented by NaN values, for certain days. The data is time-stamped, hence allows to track changes in depression levels over the course of 365 days.

- **Data Types:** The PHQ-9 scores are represented as integer values, falling within the range from 0 to 27. The patient IDs and day numbers are typically stored as alphanumeric characters.

Objective :

- The primary objective of this analysis is to detect significant shift/change points in the aggregated PHQ-9 scores, represented by CV, over the 365-day time frame.
- A change point represents a critical time point at which a substantial shift occurs in the overall depression levels of the patient population. By identifying such change points, one can aim to uncover potential patterns, trends, or triggers that may be indicative of underlying factors affecting mental health.

Approach :

To achieve the objective, the following steps have been followed :

1) **Preprocessing / Aggregation :**

At this step, an aggregated measure needs to be calculated, that should summarize the raw data, as well as carry all the inherent characteristics and distributional properties of the data.

After careful consideration and Exploratory analysis, the aggregated measure chosen for this analysis is **Coefficient of Variation (CV)**.

The coefficient of variation (CV) is a statistical measure that quantifies the relative variability of a dataset compared to its Mean / Average. It is calculated by dividing the standard deviation of the data by its mean.

A higher CV value suggests greater variation or fluctuations in the scores relative to their mean; while lower CV values indicate less variation.

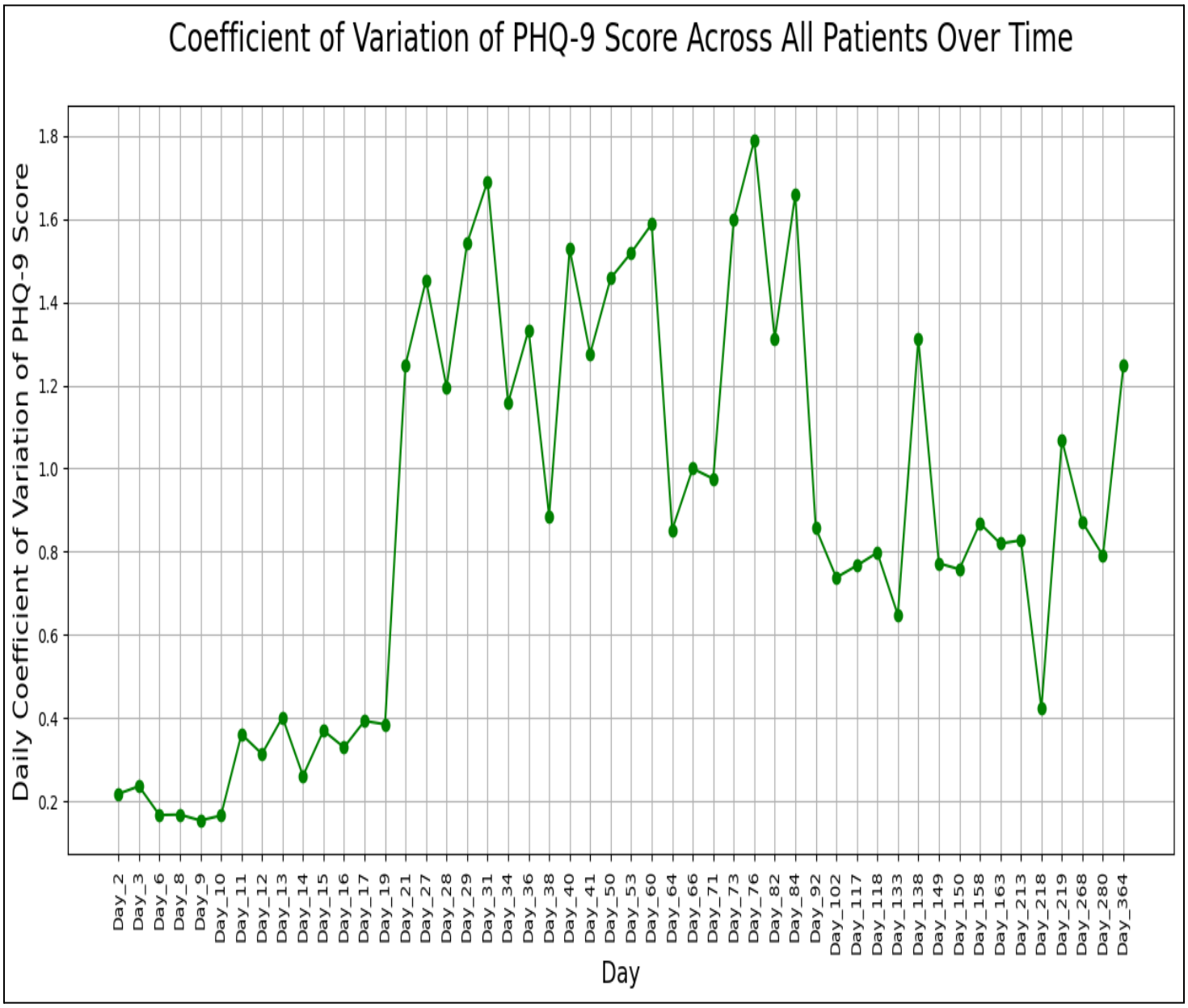
Hence, the Coefficient of Variation for each day across all the patients who have attempted the PHQ-9 survey on that day have been calculated and saved as a new series with day indices.

Aggregation Result :

As our objective is to obtain the change points along the time-axis; we may summarize and aggregate the dataset on the daily level, across all the patients who have attempted the PHQ-9 survey on that day.

So, after preprocessing, we are getting a series of Daily Coefficient of Variation with their corresponding observation Day No. as index.

After preprocessing on the sample dataset, we got a series of CV which has been visualized via Line Plot as follows:



Plot Interpretation :

From the above line plot of daily Coefficient of Variations (CV) across all patients, the following things could be interpreted :

a) At the starting days from **Day-2 to Day-10**, the mean-variance trade-off is varying within a small range & lies somewhat within a similar level.

b) After Day-10, from **Day-11 to Day-19**, there is an upward shift in the CV values, and they are staying within a small range only.

c) After Day-19, from **Day-21 to Day-60**, again an upward shift has happened in CV values and they are fluctuating within a rather bigger range over the days, but there's also a tendency could be found: which is their pattern of moving towards the mid-point of the range over the time. Here, it's clearly visible that there's a big drop at Day-38, which could be a potential change point. But also it has to be noticed that again after Day-38, the values are clustering around the average of that period.

d) From **Day-64 to Day-71**, again there's a huge drop in the CV values, and for these 3 days, the values are moving within a small range, which could be looked at as a potential separate cluster.

e) From **Day-73 to Day-84**, again there is a huge pick of the CV values that can be seen clearly & within that period they are moving within a small range only.

f) From **Day-92 to Day-364**, in spite of having a big range of variation in the CV values and some pick-drop points, it could also be seen that the values within this time range are showing a pattern of clustering around the average value. So this time period could be looked at as a separate cluster from the other ones.

❖ *From the timeseries plot of daily Coefficient of Variations (CV) over the time across all patients, it could be clearly seen that CV gives a clear picture of potential change points and is able to cluster the time-axis which have the potential to have sufficient granularity & sensitivity.*

2) <u>Change Point Detection:</u>

The next step is to perform change point detection on the calculated CV values. Change point detection is a technique that helps identify points in the data where there are significant shifts or changes in the underlying patterns or distribution.

Here, **Pruned Exact Linear Time (PELT) algorithm has been used; which internally uses Least Absolute Shrinkage and Selection Operator (LASSO) model and LAD Cost Function.**

The PELT is a dynamic programming algorithm that efficiently finds the optimal segmentation of the data based on the cost function. It searches for change points that minimize the cost associated with the shifts in the score distribution patterns.

The 'L1' model, also known as the L1-norm or Least Absolute Deviations (LAD) model, is a widely used approach in change point detection that has been used here, which refers to the L1-Pelt algorithm. The L1 regularization, also known as Lasso regularization, is applied within the Pelt algorithm to encourage sparsity and promote simpler segmentations by driving some change points exactly to zero.

The Pelt algorithm analyzes the CV values and detects change points. The cost function takes into account the CV values and penalty parameter, which controls the trade-off between detecting more change points or having a simpler segmentation.

Once the change points are detected, the data could be segregated based on these change points. Each segment represents a subset of the data with similar score distribution patterns over time.

<u>Change Point Detection Result :</u>

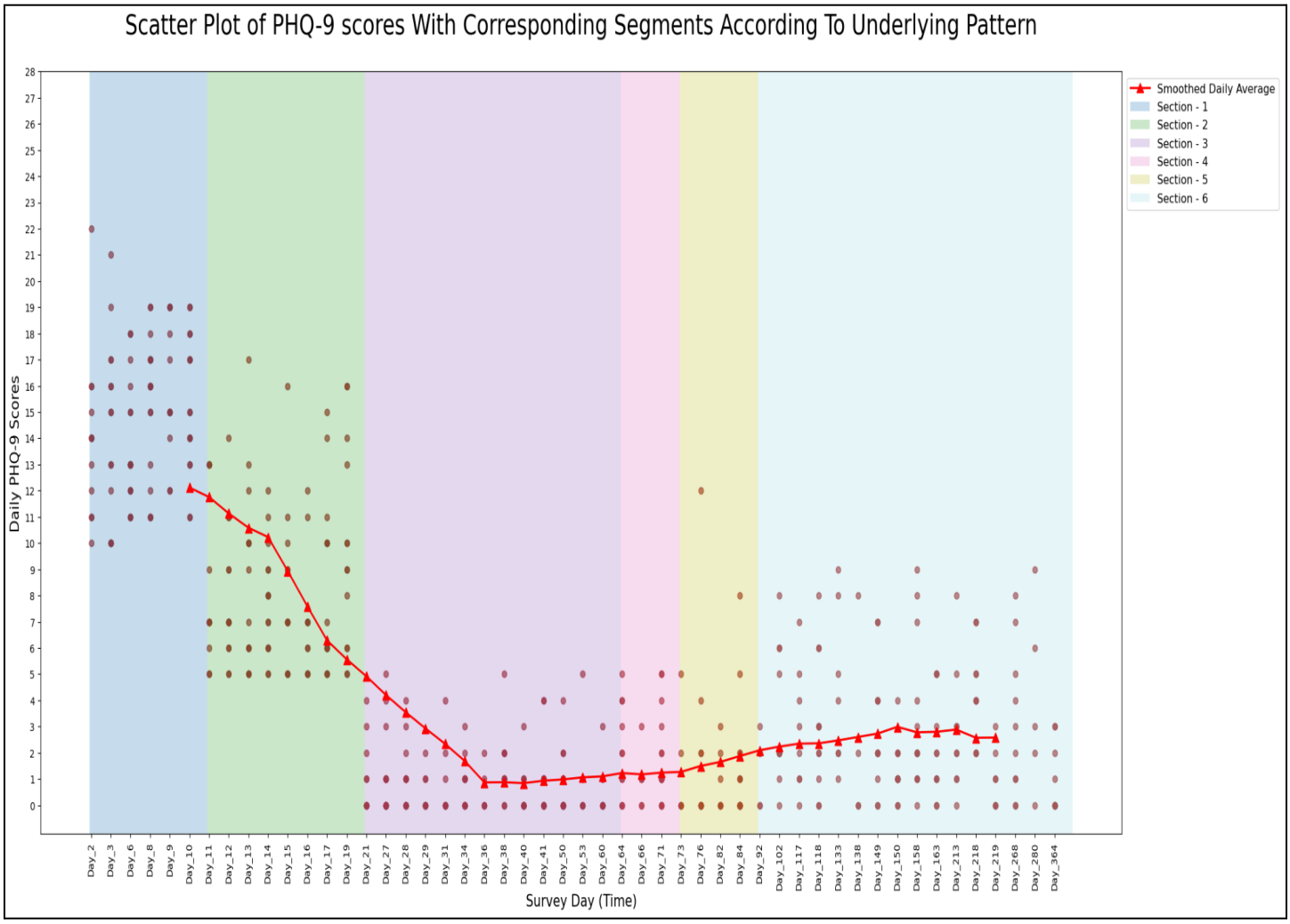
Given the sample dataset, for detecting the change points in the timeseries of daily Coefficient of Variation (calculated across all 100 patients), for sample 50 days PHQ-9 scores, the following 3 constraints have been applied on the PELT algorithm:

1) For forming a group/cluster, at least 2 points have to be considered (min_points = 2).

2) For searching change points, every point should be considered as a potential change point at the beginning (jump = 1).

3) For controlling the sensitivity of the detection, the penalty parameter has been set at the 0.5, which is the mid-point of the range of the penalty parameter (penalty can take any values within 0 to 1).

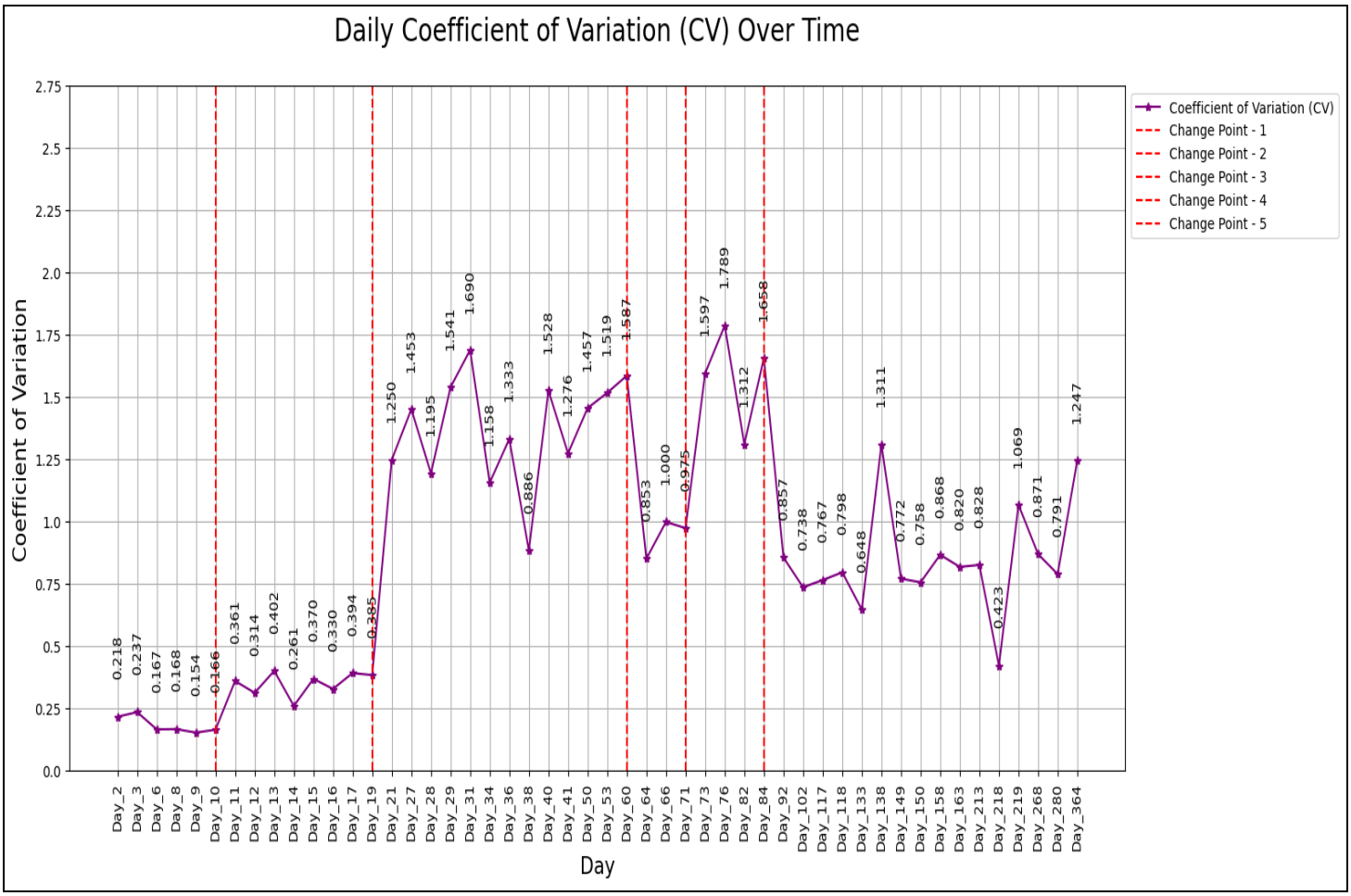
Applying these 3 constraints along with the PELT algorithm on the Daily CV scores, a total of 6 sections of the dataset have been obtained. Also, the **moving averages with a window size 10 have been calculated on the Daily Averages values across all patients to get a smoothed curve of PHQ-9 scores across all 100 patients over 50 days time span**; and hence the following result have been obtained :



3) Goodness of Fit of the Clusters :

For checking, if all the constraints that have been set for the PELT algorithm and the application of it on the timeseries of daily Coefficient of Variations (CV) across all the patients; have been satisfied and how good they have been able to cluster the time-axis based on the raw sparse matrix of PHQ-9 scores of 100 patients for 50 discrete days, lets plot the change points as vertical dotted lines of red color and along y-axis plot the daily CV scores of 100 patients and along x-axis plot the 50 discrete days for which data has been delivered.

Using the above mentioned algorithm 5 change points have been detected and in the below plot those change points have been plotted as red dotted vertical lines along with the line plot of Daily Coefficient of Variations across all patients. For getting a more clearer understanding, the CV values have been plotted as texts with their corresponding y-axis.



From the above plot, following things could be interpreted :

1) From Day-2 to Day-10, the values are exhibiting the properties :

min = 0.154,
max = 0.237,
mean = 0.185,
SD = 0.034

- It could be seen that the values are clustering around the mean and varying within a small range.

2) After Day-10, there's a shift in the CV value at Day-11 and the values after Day-11 up to Day-19, exhibiting the same pattern as well :

min = 0.261,
max = 0.402,
mean = 0.348,
SD = 0.049

- Here also, it could be seen that within this interval, the values are clustering around the mean with a small amount of dispersion.
- Also, the mean and SD of the two clusters have significant differences.
- So, it could be inferred that the clusters are well separated by the algorithm.

3) After Day-19, from Day-21 up to Day-60, there is another cluster found by the algorithm. Within this interval of time, the points are distributed with the following characteristics:

min = 0.385,
max = 1.690,
mean = 1.304,
SD = 0.338

- At this cluster, the within cluster Variance is higher that previous two clusters, there's a significant shift in the average of the CV values; also they are lying within a bigger range of dispersion.

- It could be also seen that, at Day-38, there's a huge drop in the CV value. But after that next values are clustering around the mean of the cluster values. Which is the reason this point has not been considered as a change point. Because, if algorithm mark it as a change point, then the next point i.e. Day-40 would again be a change point due to a huge pick in the magnitude. As a constraint of **min_points = 2** has been employed, solely the Day-38 cannot form a separate cluster. Which gives the proof of satisfying the **minimum point for forming a cluster**.

4) On the Day-64, there's a huge drop-off in the magnitude in the CV value and that continues up to Day-71, after which again a pick at the CV values could be seen from the line plot of CV; which indicates a separate cluster formed with the 3 points: Day-64, Day-66 and Day-71. The algorithm has been able to detect this change in the CV values as well and formed a separate cluster with these 3 points. This cluster is exhibiting the following properties :

min = 0.853,
max = 1.000,
mean = 0.943,
SD = 0.079

- Which gives a clear picture why these points are belonging to same cluster : the values of the cluster are always lying around the mean & within cluster variation is smaller.

5) After that again on Day-73, there's a huge jump in the values of CV, from 0.975 to 1.597. The algorithm has been able to detect this change and hence formed a new cluster taking the Day-73 to Day-84 as they are exhibiting the same kind of properties :

min = 1.312,
max = 1.789,
mean = 1.589,
SD = 0.201

- From the above properties and values of the points, it's clearly visible that out of 4 points of the cluster, 3 are clustering around the mean value, whereas the CV on the Day-82 is showing a drop-off. But as per the **minimum point for forming a cluster** constraint is not allowing the algorithm to form a different cluster rather include it within the cluster in which the previous and next points are exhibiting the same kind of properties. Which is also a proof the **sensitivity / penalty** of the model is medium : that means the algorithm would declare a point as a change point only when there's a moderate yet significant change in the values.

6) Finally, from Day-92 to Day-364, these points are forming another separate cluster, with the following properties :

min = 0.423,
max = 1.311,
mean = 0.848,
SD = 0.214

- Again, within this cluster, two sudden drop-off and pick-up points have been found : Pick at Day-138 and Drop at Day-218. But, again maintaining the **minimum point for forming a cluster** constraint, the algorithm has not considered these sudden pick-up or drop-off points as change points, rather checked if the next points are exhibiting the same kind of distribution just previous that sudden change point or not. If yes, **it has forcefully put that sudden change point within the cluster, where previous and next points are showing the same kind of distributional properties**.

● Conclusion :

From the above results and plots, it has been found that all the 3 constraints have been followed by the PELT algorithm for the change points detection on the time-axis in PHQ-9 scores of 100 patients

a) The minimum number of points belonging in a cluster > 2 has always been satisfied.

b) For searching the potential change points, all the points have been considered and checked as sub-problems and after checking all of them the optimal points only, which also satisfies the condition of minimization of the Cost Function, has been saved.

c) The sensitivity condition that has been set at the medium = 0.5, has been maintained too & hence the number of clusters are not too much (Over fitting), neither too less (Underfitting). Rather the algorithm has only detected those change points at which there is a statistically significant / drastic change in the distribution of the CV, compared to the neighboring points, hence the PHQ-9 scores as well (as CV is only the aggregated / summarized representation of the PHQ-9 scores, main objective was to segregate the PHQ-9 scores itself).

Finally after all the analysis and model fitting through the calculations and visualizations, we got 6 Clusters, i.e. there are 6 different sections of the PHQ-9 scores across all the 100 patients, which are significantly different from each other along the x-axis (Days); which means the sample PHQ-9 scores of the 100 patients within the 50 sample discrete days are showing different inherent/underlying distributions of scores over the time, which shows a clear picture of the effect of treatments by a particular clinic/organization on these group of patients in their Depression Levels from their admission for the treatment to recovery period up to 1 year.