Introduction - Purpose:

The schools run various remedial program for heping out students who have done poorly in assessments.

The logic behind selection for students for such programs often varie from school to school is very dependant on the teacher.

Clustering the students on the basis of their past assessment results will bring out patterns that will help us classify them into natural groups consisting of students of similar level of competency

Currently, we have assessment data for various FA, SA tests. The idea is to look at two factors for clustering :

- Their overall 'average' performance across the assessments
- The variability of their scores across the assessments

The idea is to be able to treat student who have been consistently doing poorly/well different from the student who are more irregular with both high and low scores.

Toggling the raw codes:

```
Out[1]:
```

Click here to toggle on/off the raw code.

Importing Libraries and connecting to SQL:

Connecting to SQL server

Enter password to connect to the Samarth Staging server :

```
........
```

```
Connected successfully
```

Clustering example:

Subsetting for Shimla

As I cannot pull the entirey of the data, I am currently using only data from the Shimla district. A sample can be seen below:

```
Out[33]:
```

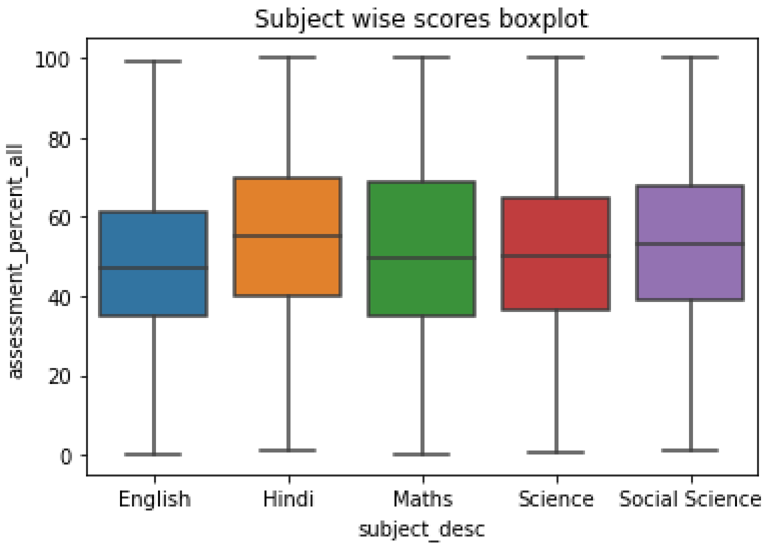
	assessment_percent	assessment_grade	assessment_grade_all	assessment_percent_all	su
0	-1.0	A	A	87.5	
1	-1.0	A	A	87.5	
2	-1.0	A	A	87.5	
3	-1.0	A	A	87.5	

4 rows × 25 columns

We have filtered for students of 9th and 10th grade. Looking at the distribution of scores for each subject :

```
Out[39]:

Text(0.5, 1.0, 'Subject wise scores boxplot')
```



We see that Hindi registers the highest average score and Math the biggest spread.
On the other hand, English shows the lowest average scores.

Understanding the assessment percentage and percentile metrics :

At a student assessment level we have measured their performance in two ways :

- Assessment percentage : This is the assessment score expressed in terms of percentage to have consistency across various assessments
- Assessment percentile : This is the assessment score expressed in terms of percentile comparing the score to all the other students of that district for that particular assessment

Example of difference in average test performance across assessments:

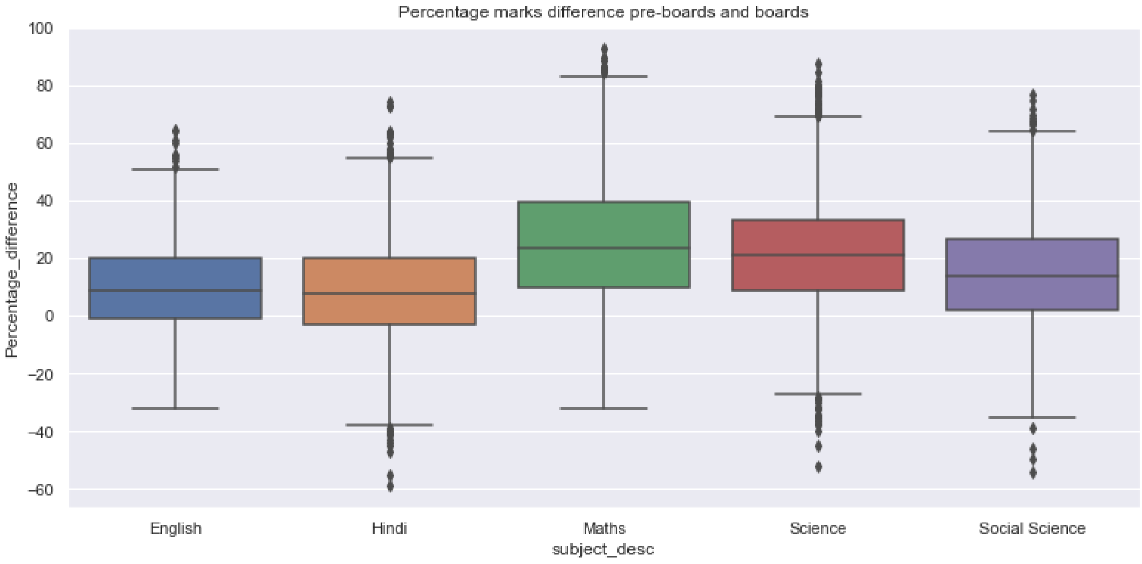
Looking at the assessment percentage differences across tests:

It must be pointed out that the assessment percentage might not necessarily be a good indicator of a students performance across tests relative to other students.

For example if we look at the difference between the board exam and the pre-board exam results for students, we can see clearly see that there is an average increase in the scores of each student. For Maths/Science there is an average increase of the score by more than 20, while in other subjects it's around 10

```
Out[245]:
```

```
Text(0.5, 1.0, 'Percentage marks difference pre-boards and boards')
```

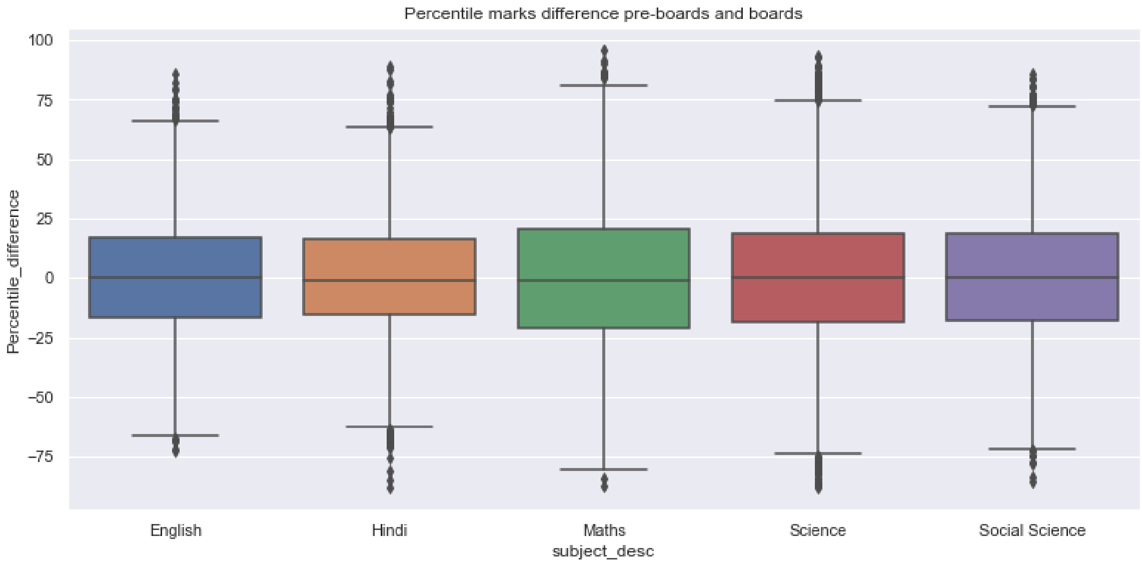


Looking at the difference in percentile scores:

When we look at the difference in the percentile scores, we see however that in general there is an average of 0 difference in the percentile scores between the two exams

```
Out[246]:
```

```
Text(0.5, 1.0, 'Percentile marks difference pre-boards and boards')
```



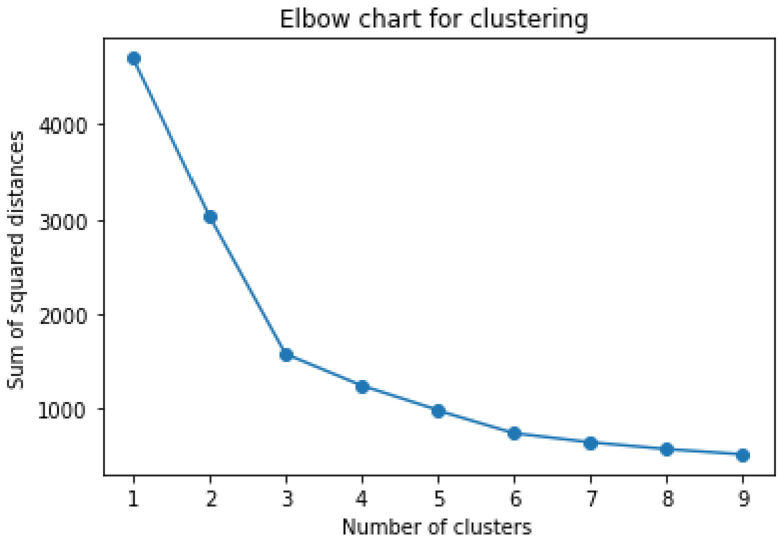
This means that an average student could get around 40 in the pre-boards in Math and around 60 in the boards. His percentile score will remain the same at around 50

Also, we can see here that Math has the biggest spread i.e. shows the most changes in percentile scores. This is further proof of it being the subject where one can see maximum change in performance scores with some kind of intervention

Create clusters

Elbow chart

The elbow chart is used to get an idea on what are the number of clusters to be created such that the patterns are separate from each other. The number at which the elbow is being formed is the point at which sharp clusters are being defined and should be used



We see that 3-4 is the number of clusters for which an elbow is being formed. We decide to go with 4 clusters

Creating clusters across the 5 subjects for the students (currenlty in 10th grade):

Based on the earier decided logic, we need to select 2 metrics to get a mesure for average performance of the student across tests and the in-consistency in their performance:

Average assessment percentage:

We decided to take a simple average of the percentage scored in the assessments as the first measure for clustering

At a test level, the percentile cannot be considered as they form a uniform distribution and the percentile is defined such that they don't get clustered.

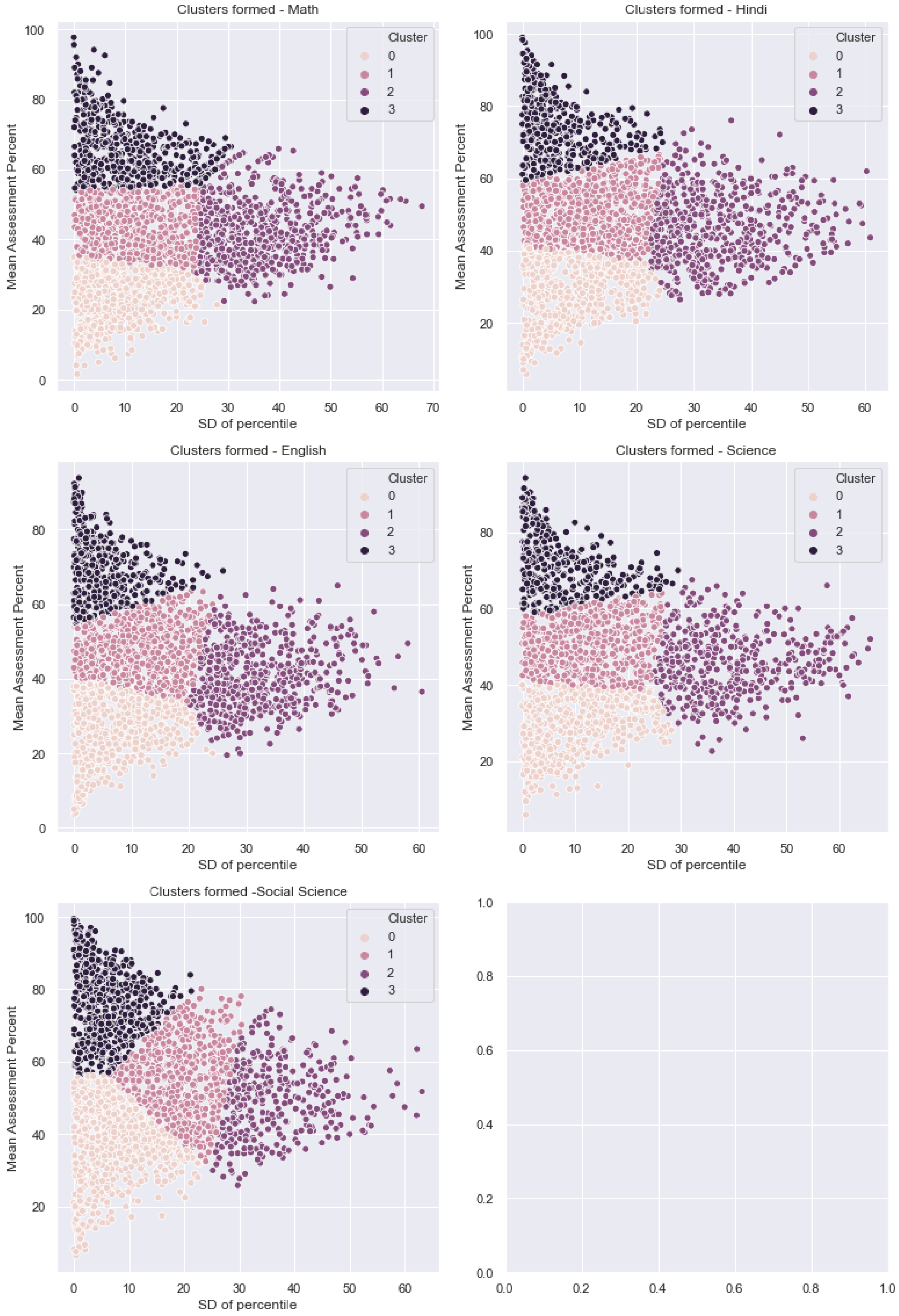
Standard deviation of the assessment percentiles:

For capturing the variation in the performance across tests, we use the standard deviation of the percentile as the metric

As shown earlier, there is often variation between the assessment percentage across tests even if relative performance of the student remains unchanged. Thhis is mainly due to relative change in the difficulty of the exams, varying importance of the exams etc

Hence, to correctly capture the variation in performace of a students compared to other students in a cohort, the variance of the percentile is a more appropriate metric

Carrying out the clustering process:


Clusters formed - Math
Mean Assessment Percent
SD of percentile
0
20
40
60
80
100
0
10
20
30
40
50
60
70
Cluster
0
1
2
3
Clusters formed - Hindi
Mean Assessment Percent
SD of percentile
20
40
60
80
100
0
10
20
30
40
50
60
Cluster
0
1
2
3
Clusters formed - English
Mean Assessment Percent
SD of percentile
0
20
40
60
80
0
10
20
30
40
50
60
Cluster
0
1
2
3
Clusters formed - Science
Mean Assessment Percent
SD of percentile
20
40
60
80
0
10
20
30
40
50
60
Cluster
0
1
2
3
Clusters formed -Social Science
Mean Assessment Percent
SD of percentile
20
40
60
80
100
0
10
20
30
40
50
60
Cluster
0
1
2
3
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

We see that 4 similar clusters are being formed across the subjects. The clustering process was modified to ensure that similar clusters are simlarly labelled across the subjects.

For example,the bottom cluster is always labelled as 0 across the subjects

Cluster Interpretation:

Cluster 0 :

This is the cluster with the lowest mean score and lowest standard-deviation of percentile (bottom left cluster)

These are the students who are consistenly doing poorly.

These student must be the focus of the basic programs as they are consistently failing the assessments and are behind the rest of the class in understanding the outcomes.

Cluster 1 :

This is the cluster with the medium mean score and low standard-deviation of percentile (bottom left cluster)

These are the students who have an average score and haven't shown much variance in their scores. The lower half must be watched carefully so that they don't slip into the poorly perfroming cohort.

Cluster 2 :

This is the cluster with the medium mean score and high standard-deviation of percentile (bottom left cluster)

These are the students who have an average score but have at some point shown good performance. THese students have good potential and must be the focus of programs that look at improving the school merit performance as they could be pushed more easily to get higher scores

Cluster 3 :

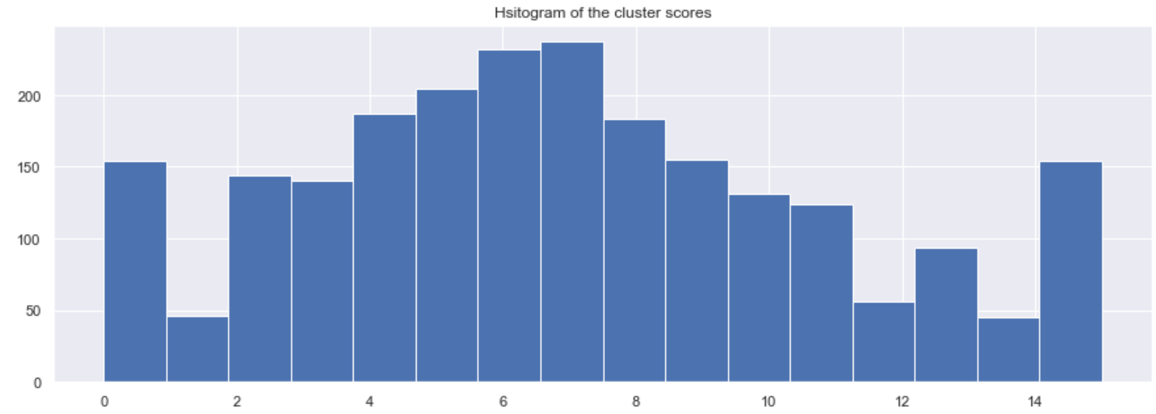
This is the cluster with the high mean score and low standard-deviation of percentile (bottom left cluster)

These are the students who have consistenly shown good performance and are the least cause of worry

Looking at students belonging to clusters across subjects:

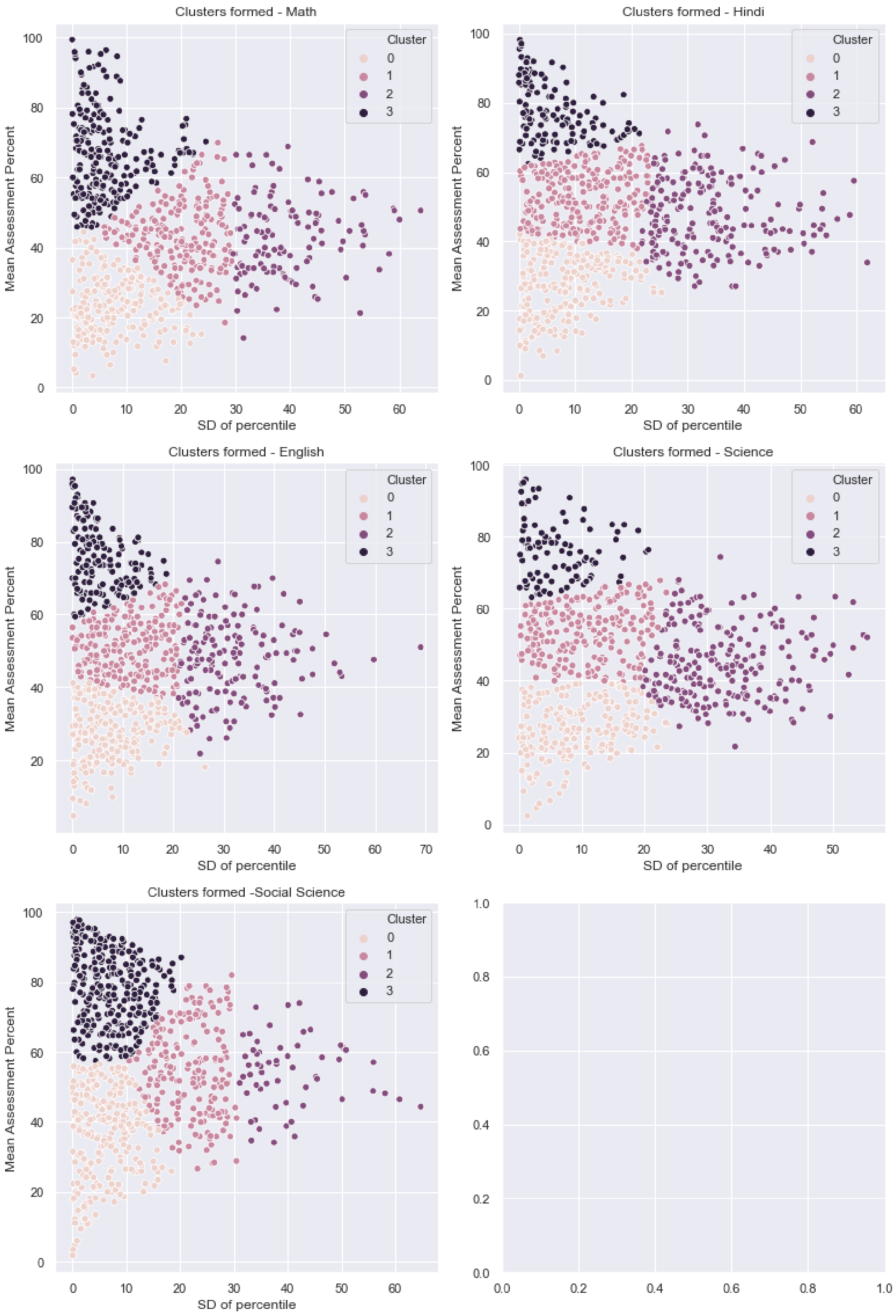
At a student level, I have summed up the clusters assigned to a student for each subject. For example, if a student belongs to cluster 0 in all the 5 subjects, his total cluster score will be 0 If they belong to the top cluster (3) in all the 5 subjects, their cluster score will be 15

Plotting ahistogram of the same:



We see that is forms the usual normal curve except at the edges, the students who belong to cluster 0 in one subject tend to belong to cluster 0 in all of them and those who belong to 3 in one, tend to belong to 3 in all of them

Re-doing for Hamirpur:



We can see similar clusters being formed with less data points for a more backward district

Potential pain point on expanding: The cluster shape may vary a bit on adding more assessments. It is likely that the average variance will decrease on adding more tests. However we still expect 4 clusters to be formed and th logic to hold, the difference is that the differences in variance between the clusters will be lower