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 Prod server :

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

```
Connected successfully
```

User defined functions for clustering:

Creating the datasets:

Pulling from SQL :

The required dataset is pulled from the prod Samarth DB.
The data is being pulled for Grade 7 and for Shimla district. They have data only for FA1,FA2,SA1 and SA2.
The SQL code is based on the EDA done for new dataset-Samarth DB.

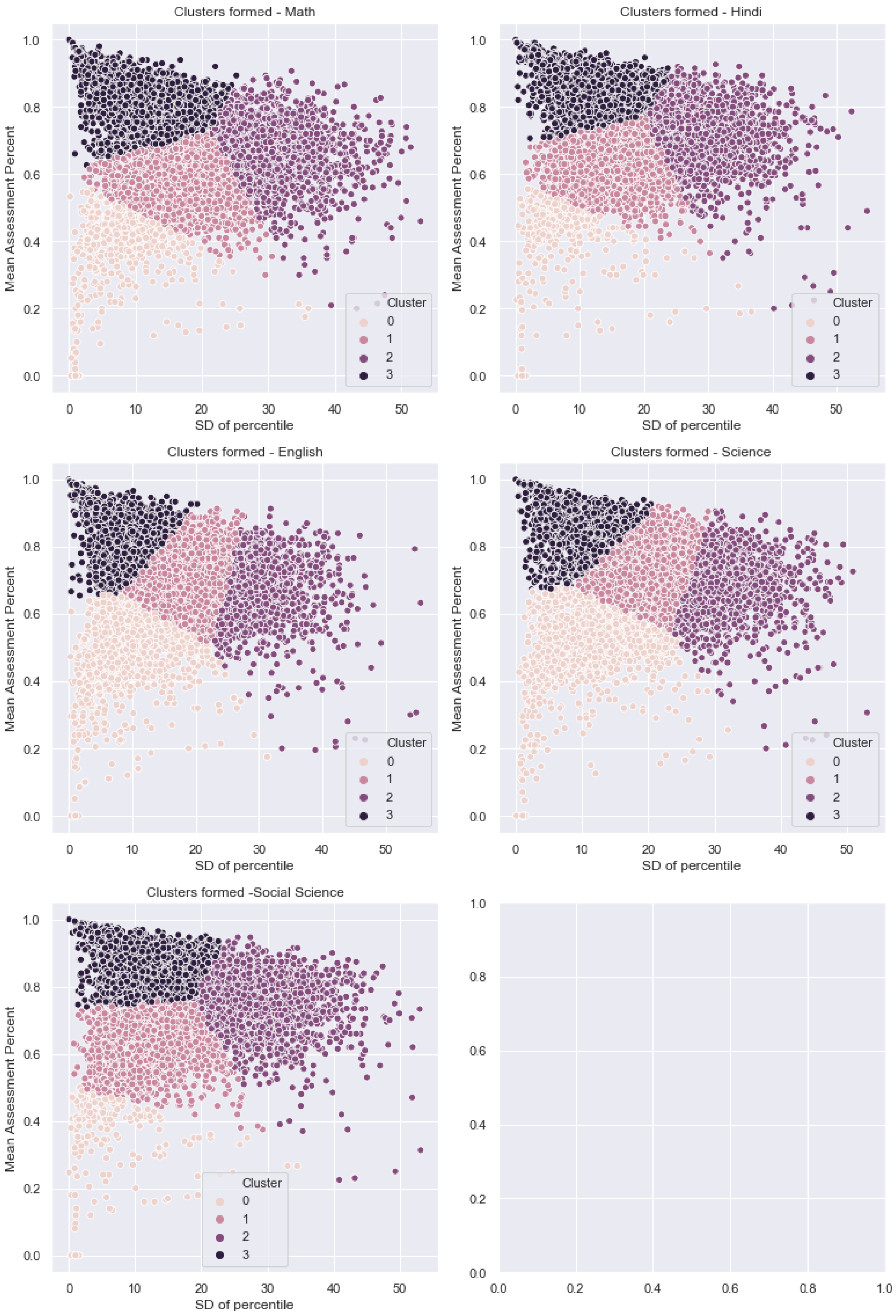
Filtering on the students:

- Only considering students who have written at least 2 assessments

Preparing clustering data

- Calculating the mean percentage and the standard deviation of the percentile. The clustering method is discussed in 'Clustering Use Case- Samarth' data notebook

Carrying out clustering and creating metrics


Clusters formed - Math

Clusters formed - Hindi

Clusters formed - English

Clusters formed - Science

Clusters formed -Social Science

Collating clustering data and combining with other student info :

Creating student level cluster subject combinations :

For each cluster, I've calculated the combinations of subject for each student belongs to that cluster.

For example, If a student belongs to the top cluster (cluster 3) in all the subjects, then for cluster 3 it should show 'English, Maths, Social Science, Hindi, Science' against him but if he belongs to the top cluster onnly for Engish and Hindi and not for other subjects, then it should show 'English,Hindi' against him.

This will be useful in creating separate groups of students for the remedial classess/ speical programs and holding classes for them parallely.
For example, if there are 10 students in cluster 0 across all subjects and 6 of them are in cluster 0 for Hindi, English and 4 of them in Science, Maths. Then one can hold parallell sessions for these students

Adding LO information to the students:

Each assessment is connected to LO buckets which are in turn connected to multiple LO's. We can calculate the scores for each of the LOs on this basis and over many assessments calculate what is the worst perfroming and best performing LOs

Saving the dataset :