Mid Evals - SMAI

Team Number: 55 | Team Bash Party

- Keshav Bajaj (2019115010)
- Tejasvi Chebrolu (2019114005)
- Naman Ahuja (2019101042)
- B Vaibhaw Kumar (2019112021)

Copy of Timeline

Timeline	Milestones	Status
@November 1, 2021 → November 7, 2021	Literautre Review	Completed
@November 7, 2021	Project Proposal Submission	Completed
@November 7, 2021 → November 10, 2021	Creation of pipeline and preparing dataset	Completed
@November 11, 2021 → November 12, 2021	Run PCA on Datasets	Completed
@November 13, 2021	Running other DRAs	Completed
@November 14, 2021 → November 16, 2021	Convert Euclidean distances into conditional probabilities that represent similarity	Facing Difficulties, Ongoing
@November 17, 2021 → November 20, 2021	Mid Project Evaluations	Completed
@November 21, 2021 → November 27, 2021	Implementing improvements and suggestions based on mid evals	TBD
@November 28, 2021 → November 29, 2021	Testing	TBD
@November 30, 2021	Write Report and make presentation	TBD
@December 1, 2021 → December 4, 2021	Final Presentation	TBD
@December 4, 2021	Final Report submission	TBD

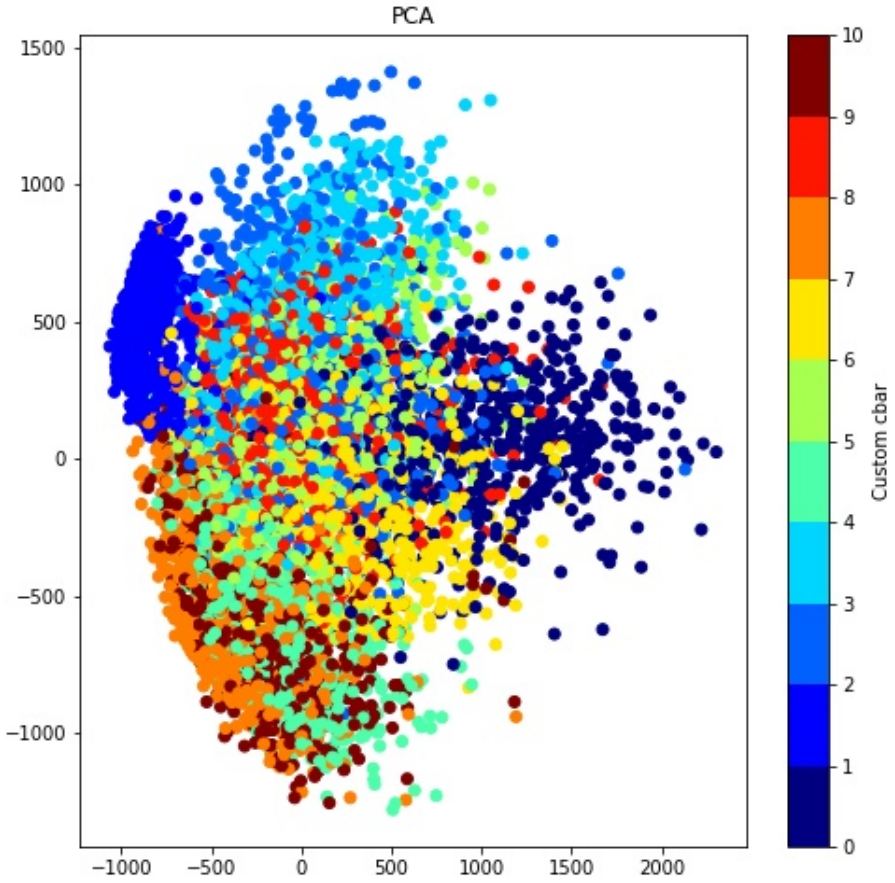
Work Done till Now

- Literature Survey: We explored and studied different dimensionality reduction techniques used for transforming and visualizing high dimensional data.
- Dimensionality Reduction Techniques: We explored and studied different dimensionality reduction techniques used for transforming and visualizing high-dimensional data.
    - Principal Component Analysis: is used to explain the variance-covariance structure of a set of variables through linear combinations. It is often used as a dimensionality reduction technique. It is the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest.
    - Sammon Mapping: **Sammon mapping** or **Sammon projection** is an algorithm that maps a high-dimensional space to a space of lower dimensionality by trying to preserve the structure of inter-point distances in high-dimensional space in the lower-dimension projection.
    - Isometric Mapping: **Isomap** is a non-linear dimensionality reduction method. It is one of several widely used low-dimensional embedding methods. Isomap is used for computing a quasi-isometric, low-dimensional embedding of a set of high-dimensional data points. The algorithm provides a simple method for estimating the intrinsic geometry of data manifold based on a rough estimate of each data point's neighbours on the manifold. Isomap is highly efficient and generally applicable to a broad range of data sources and dimensionalities.

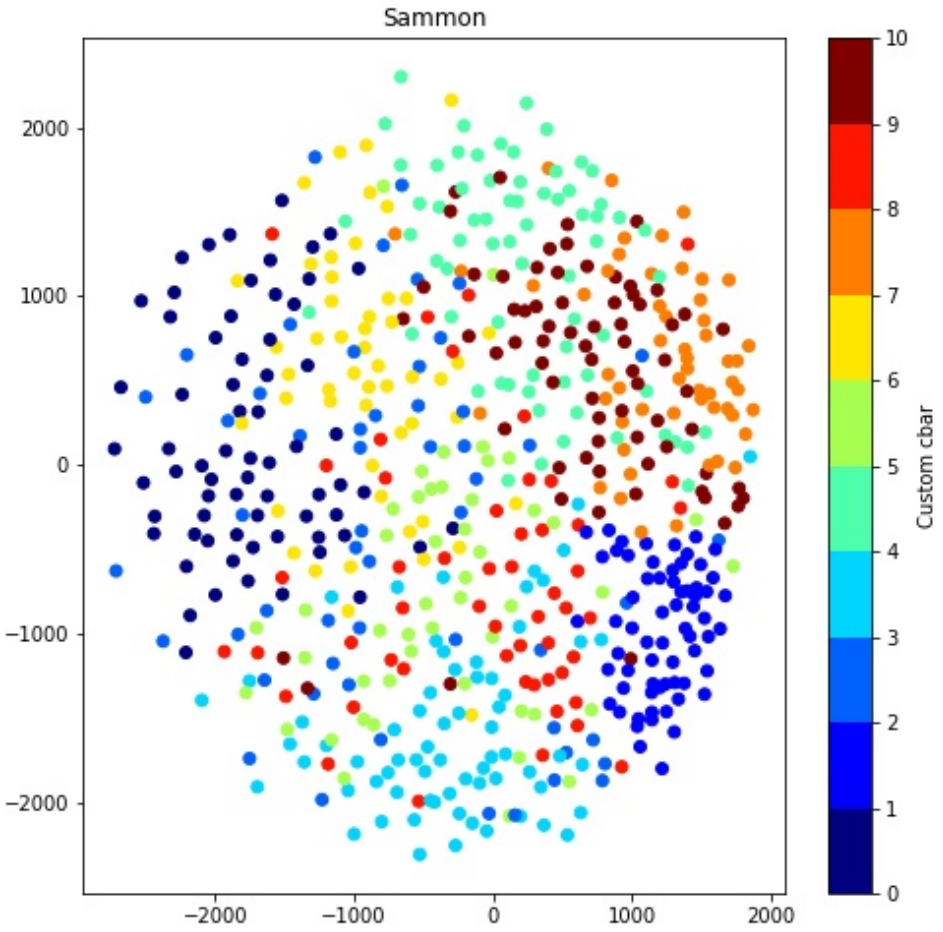
- Understanding MNIST and Olivetti Face Datasets: We used two datasets to try out the above-mentioned DR techniques to understand their working.
    - MNIST Dataset: The MNIST(Modified National Institute of Standards and Technology) database consists of handwritten digits. It has a training set of 60,000 examples and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centred in a fixed-size image.
    - Olivetti Faces Dataset: This dataset contains a set of images of faces taken between April 1992 and April 1994 at AT&T Laboratories Cambridge. There are ten different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling / not smiling), and facial details (glasses / no glasses). All the images were taken against a

dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement).

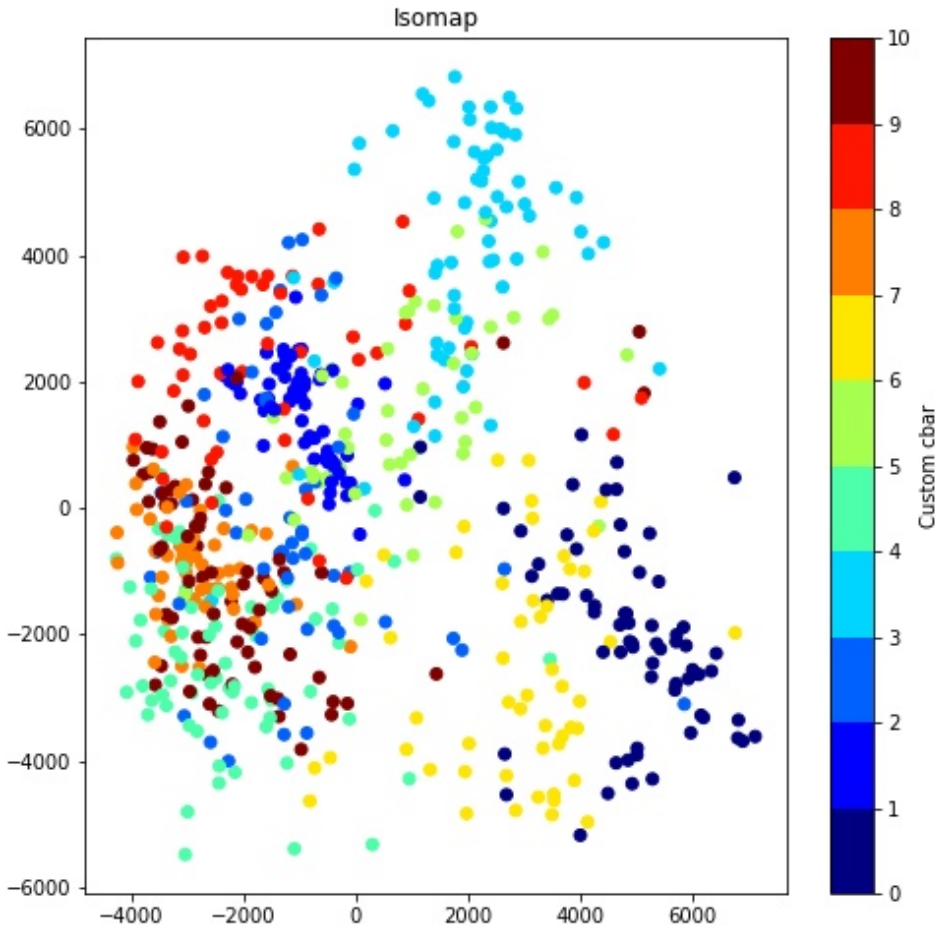
- Ran PCA, Sammon, Isometric Mapping on MNIST and Olivetti Faces datasets for dimensionality reduction.
    - MNIST and Olivetti Faces
        - Reduced the initial 784 dimensions to 2 using Principal Component Analysis (Visualization shown below)
        - MNIST Dataset (subsample of 600 points reduced to 2 dimensions using PCA)
        - Reduced the initial 784 dimensions to 30 using PCA
            - PCA after reducing dimensions from 784 to 2



- Sammon Mapping to reduce 30 dimensions to 2 for visualisation

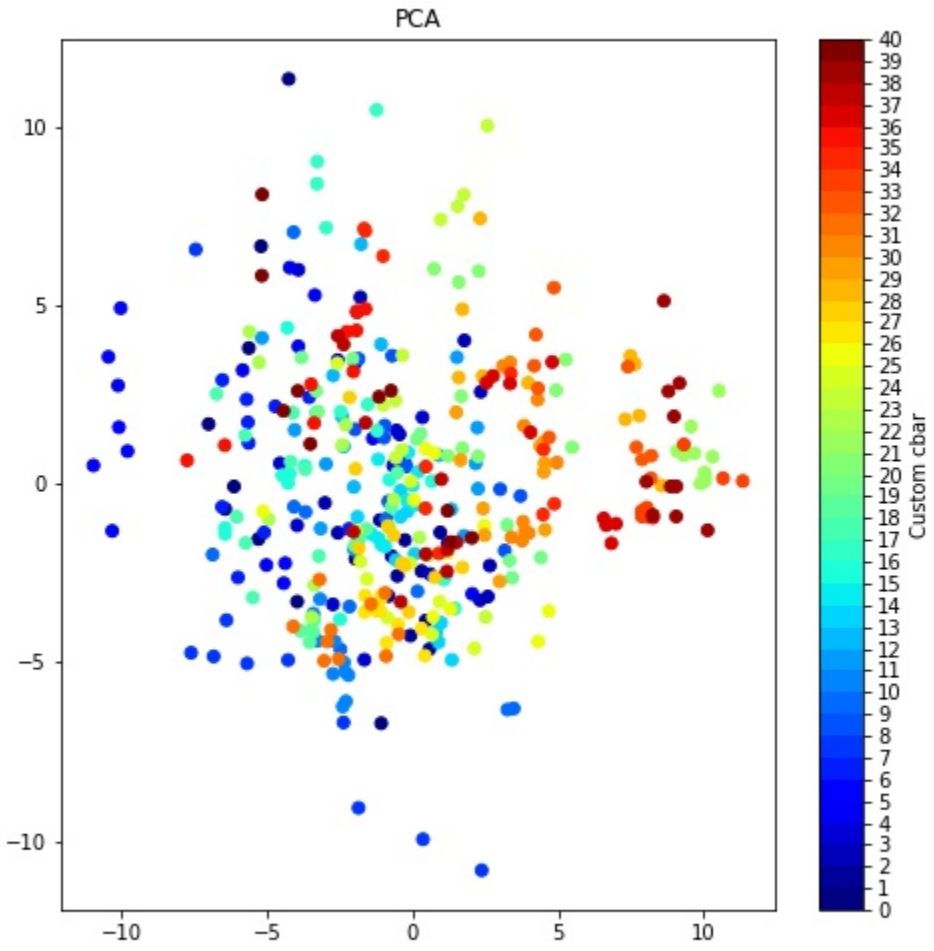


Isometric Mapping to reduce 30 dimensions to 2 for visualization

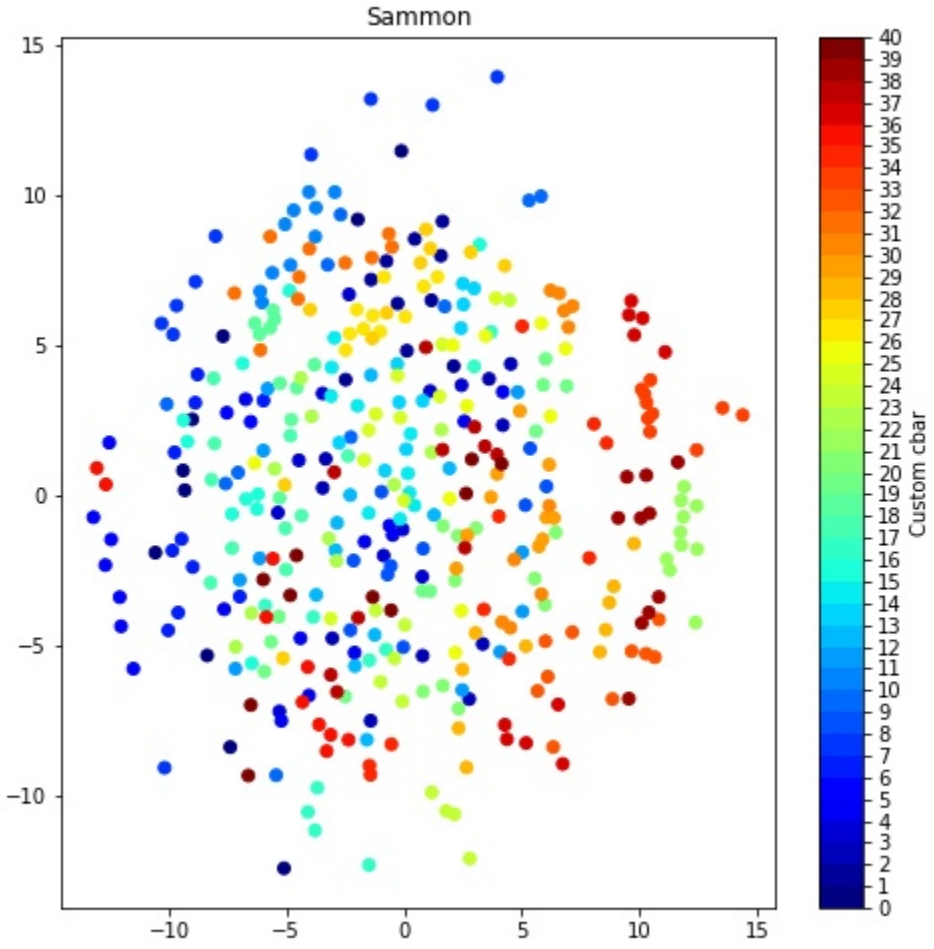


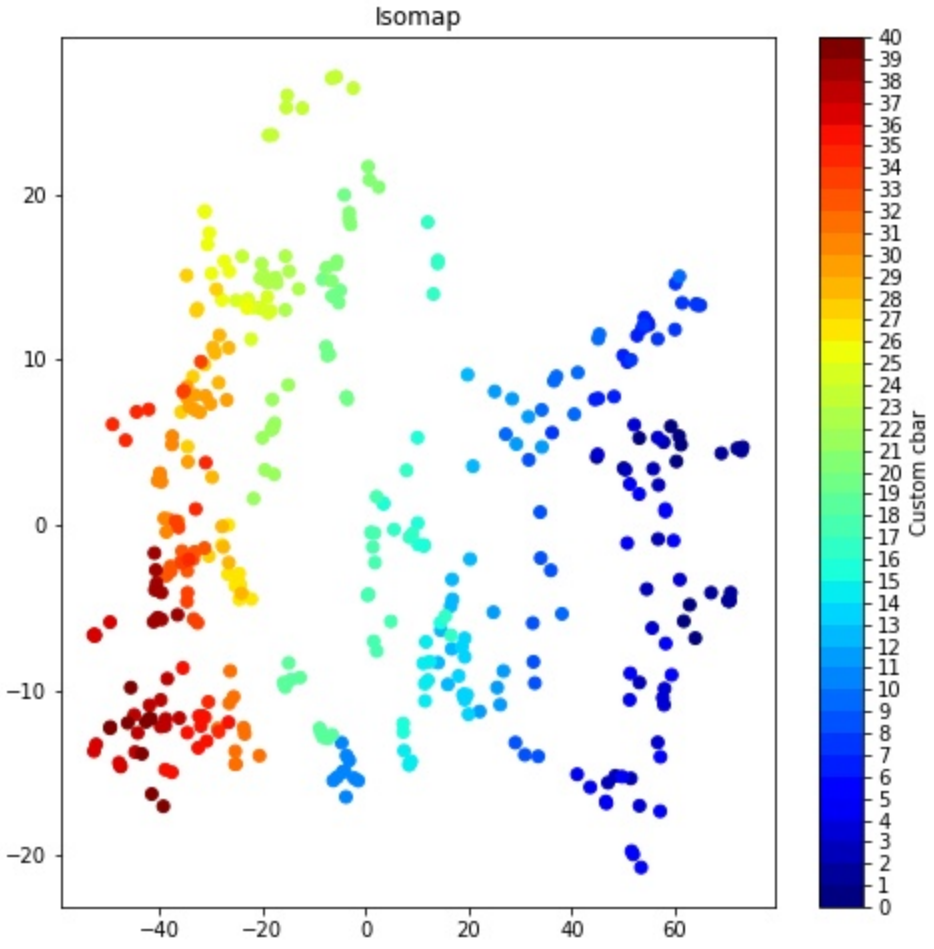
Olivetti Dataset

- Ran PCA on Olivetti Dataset to reduce 4096 dimensions to 2



- Ran PCA to reduce 4096 dimensions to 30. Then ran Sammon and isometric mapping for visualization.


Sammon
15
10
5
0
−5
−10
−10
−5
0
5
10
15
Custom cbar
40
39
38
37
36
35
34
33
32
31
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0



Future Work

To have a quantitative and qualitative understanding of the pros and cons of **t-SNE**. When it would be a good idea to use it, and when it would not be a good idea. We would have a working implementation of:

- **t-SNE:**
  - Code
  - Scatterplots for different datasets
- **Modified t-SNE:**

- Code
- Scatterplots for different datasets