CSE-564: Visualization

Mini Project #2 Report

Saurabh Sanjay Agrawal (113278626)

Project Overview:

- This project focuses on the dimensionality reduction using PCA and visualization of CSV data in the form of Scree Plot, Scatterplot Matrix, PCA Biplot, MDS plots with Euclidian and Correlation Distance and Parallel Coordinate Plot along with Elbow Plot and k-means clustering.

Data:

- The data used for this project is ‘Udemy courses for Development’ and it is obtained from following source in Kaggle: https://www.kaggle.com/jilkothari/udemy-courses-development

Attributes description:

Following are the attributes I chose from the dataset which I thought would be interesting to visualize and analyse:

- id: The course ID of that particular course.
- title: Shows the unique names of the courses available under the development category on Udemy.
- url: Gives the URL of the course.
- is_paid: Returns a boolean value displaying true if the course is paid and false if otherwise.
- num_subscribers: Shows the number of people who have subscribed that course.
- avg_rating: Shows the average rating of the course.
- avg_rating_recent: Reflects the recent changes in the average rating.
- num_reviews: Gives us an idea related to the number of ratings that a course has received.
- num_lectures: Shows the number of lectures the course offers.
- num_practice_tests: Gives an idea of the number of practice tests that a course offers.
- created: The time of creation of the course.
- published_time: Time of publishing the course.
- discount_price_amount: The discounted price which a certain course is being offered at.
- discount_price_currency: The currency corresponding to the discounted price which a certain course is being offered at.
- price_detail_amount: The original price of a particular course.

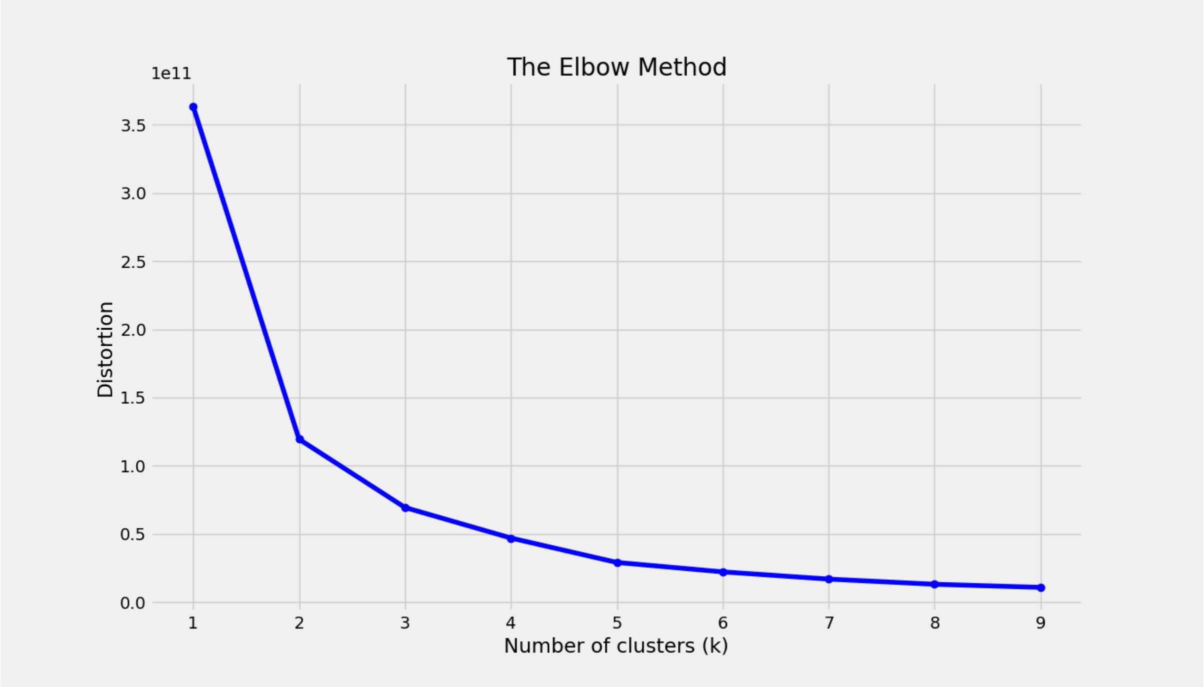
- price_detail_currency: The currency corresponding to the price detail amount for a course.

Reason for the choice of dataset:

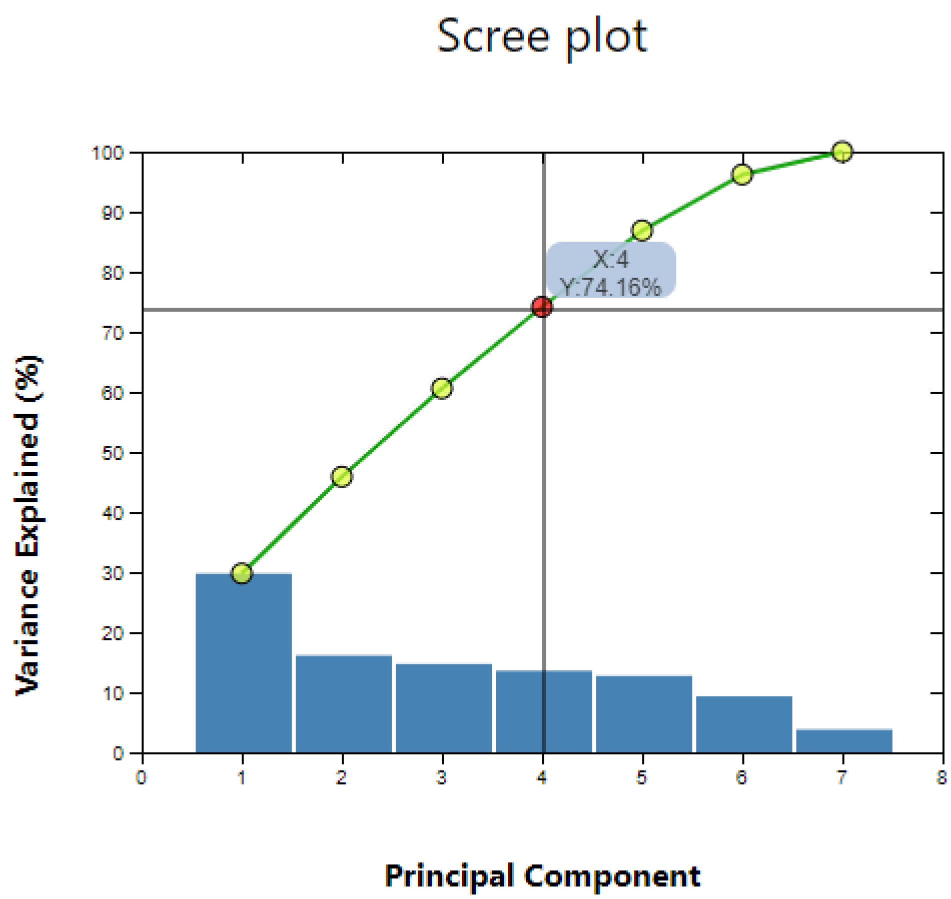
- This dataset contains a good amount of numerical features which can be of great use for dimensionality reduction using PCA.

Features implemented in this project:

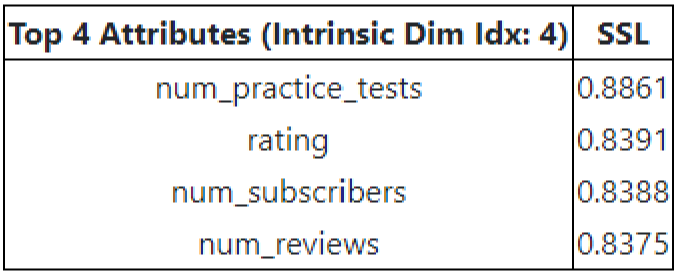
1. Elbow plot
   - Helps in finding optimal k (number of clusters) in k-means clustering algorithm
   - There's an elbow at k = 2 in the below figure



2. Scree Plot
   - Helps in displaying Principal Components as bars representing 'Variance Explained (%)' on Y-axis and Principal Components on X-axis.
   - User can select 'Intrinsic Dimensionality Index' by clicking on the circle on top of each bar, which is on the line plot showing cumulative 'Variance Explained (%)' for each Principal Component
   - This user interaction with the scree plot also makes updates in the Table showing Top 4 attributes and Scatterplot Matrix
   - On hovering on the circle, tooltip is displayed showing X and Y axis values
   - Crosshair is also implemented for good UX
   - On click of a circle, it turns red, so that user gets to know what 'Intrinsic Dimensionality Index' was selected

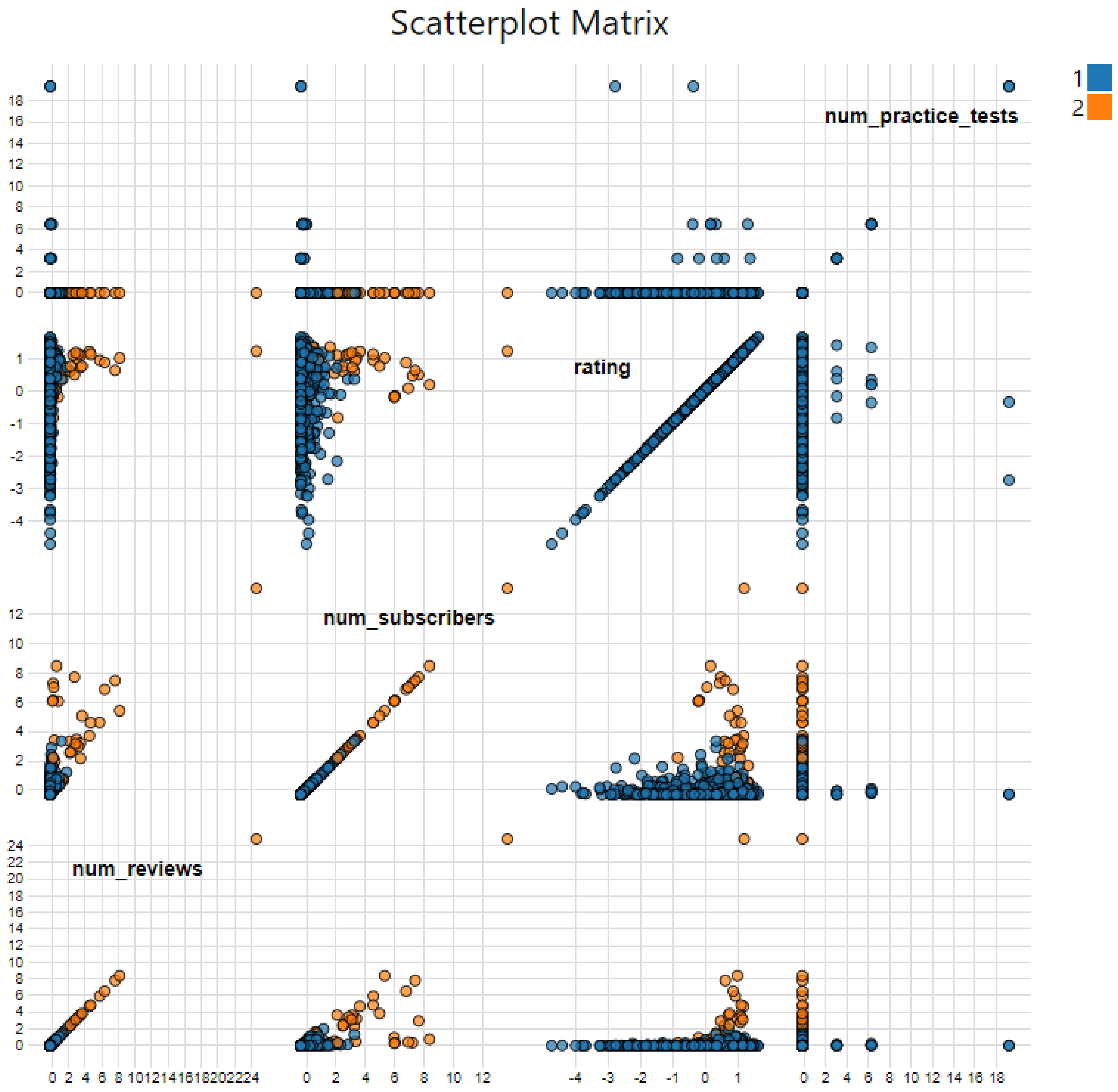


3. Table with top 4 attributes
   - This table shows top 4 Attributes and Sum of Squared Loadings value (SSL) based on the 'Intrinsic Dimensionality Index' selected from the Scree Plot.



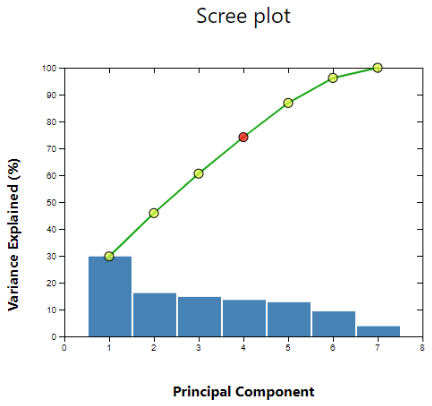
Top 4 Attributes (Intrinsic Dim Idx: 4)	SSL
num_practice_tests	0.8861
rating	0.8391
num_subscribers	0.8388
num_reviews	0.8375

4. Scatterplot Matrix
   - Helps in visualizing relationships between each combination of Top 4 Attributes in 2D in the form of Scatterplot for each combination
   - Legend on the right signifies the clusters

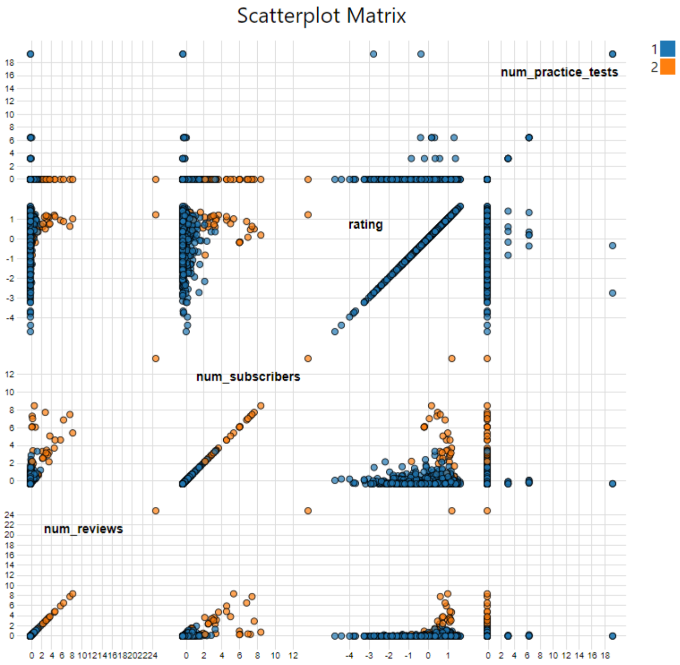


Scatterplot Matrix

○ All these three elements are kept on the same page for better analysis and UX

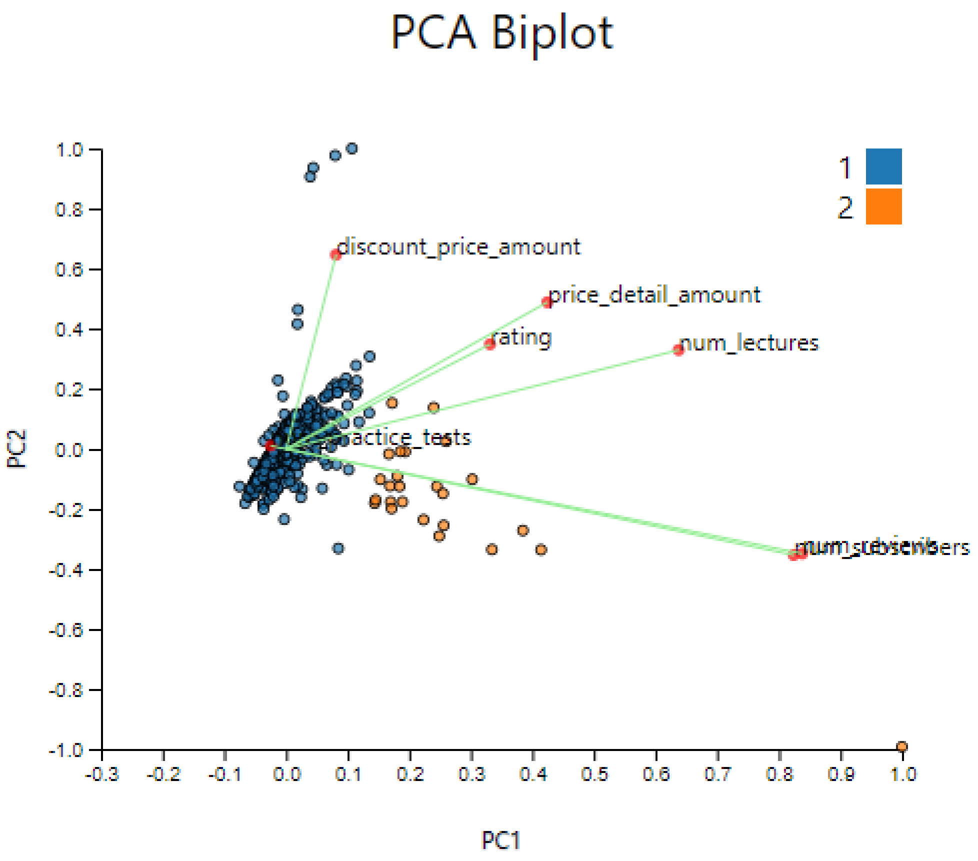


Top 4 Attributes (Intrinsic Dimensionality Index: 4)
num_practice_tests
rating
num_subscribers
num_reviews

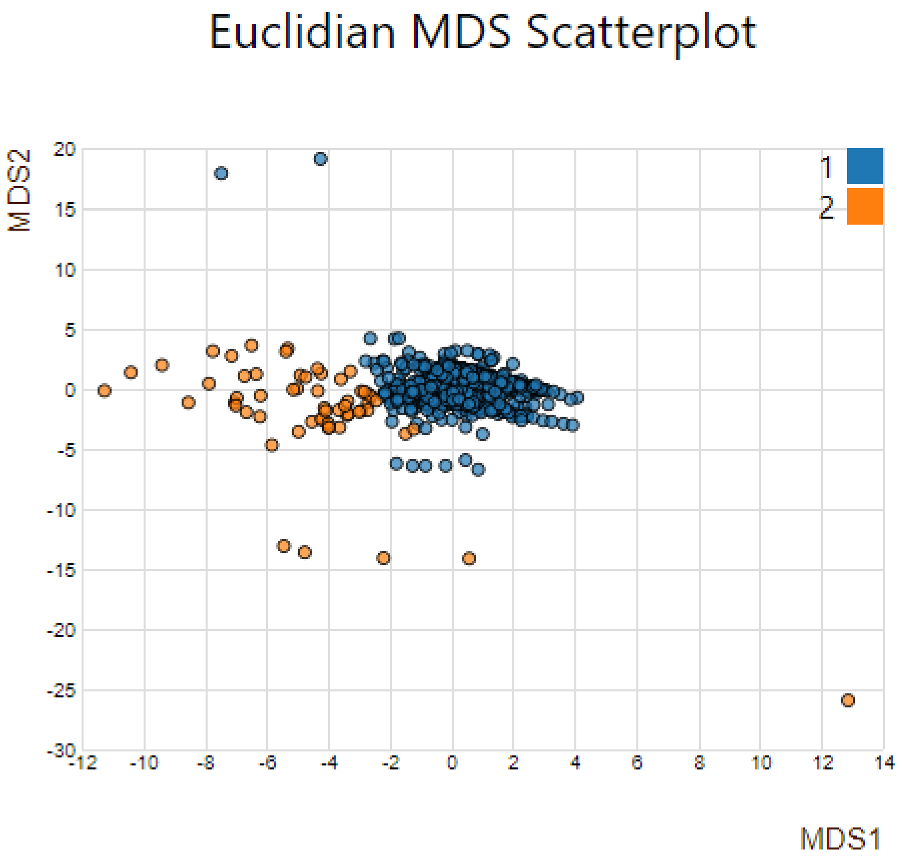


5. PCA Biplot

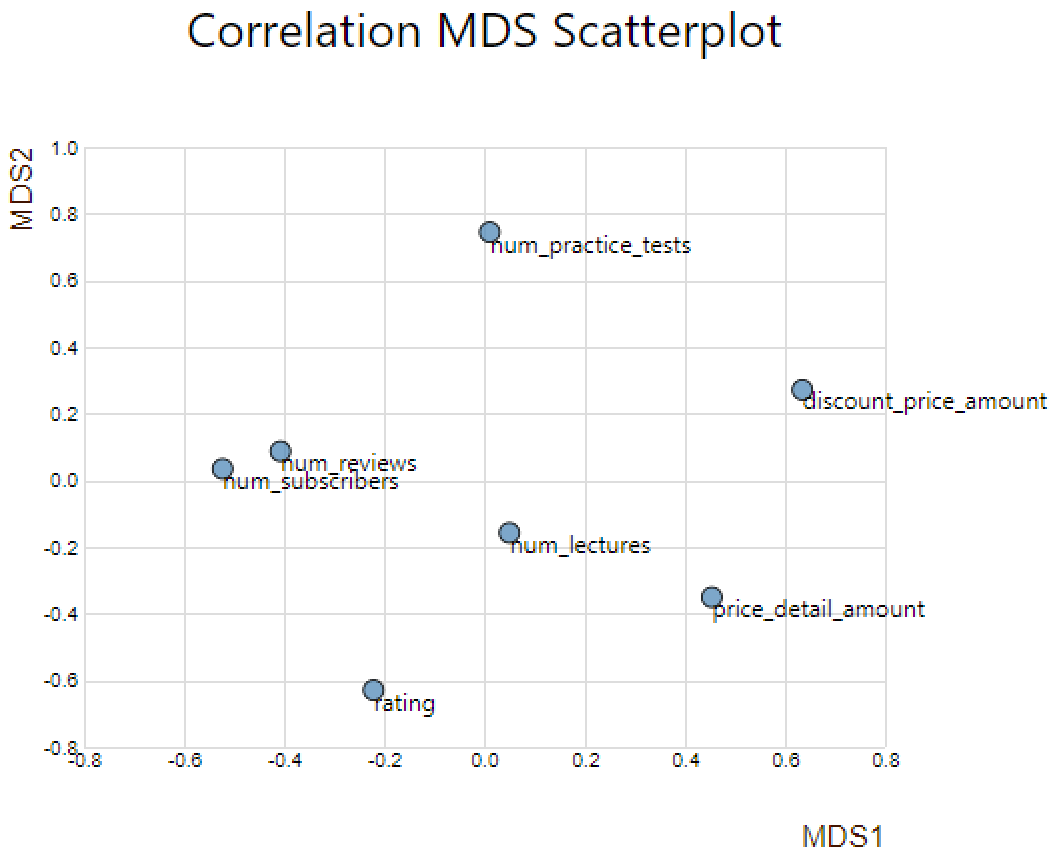
- Helps in visualizing information on both samples and variables of a data matrix for Top 2 Principal Components
- Legend on the right signifies the clusters



6. MDS plot (Euclidian distance)
    - ○ Helps in mapping the distances between observations from N-D into low D (1, 2 or 3)
    - ○ Ensure that differences between pairs of points in this reduced space match as closely as possible

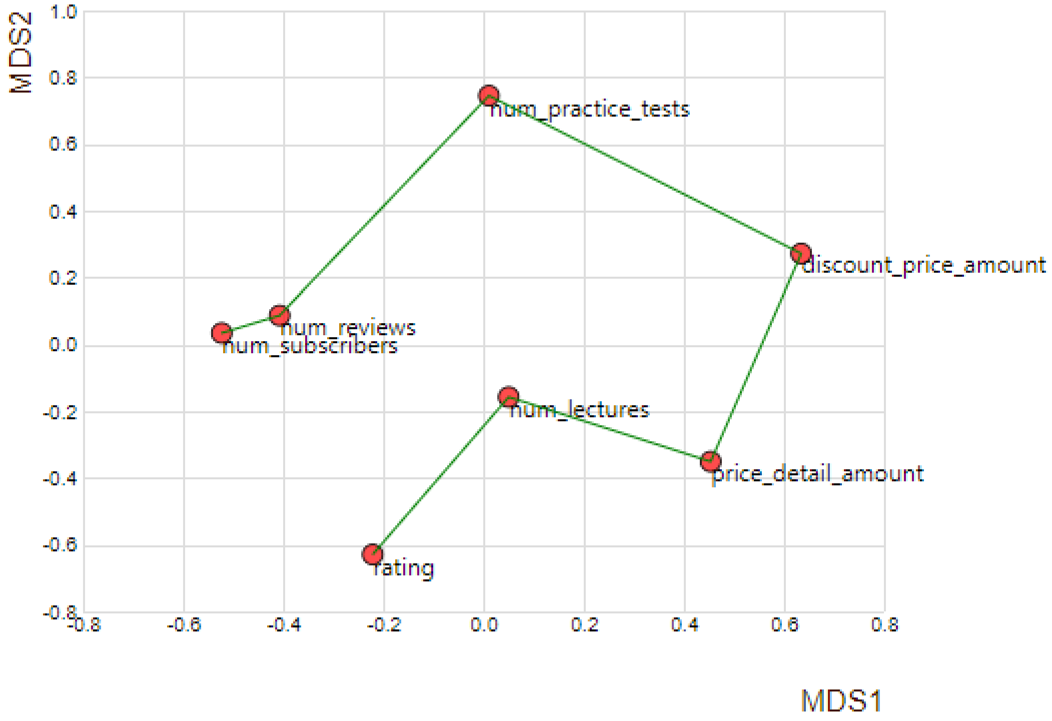

Euclidian MDS Scatterplot

7. MDS plot (Correlation distance) for variables
    - ○ Helps in visualizing the N-D variables based on ‘1 - |correlation|’ distance



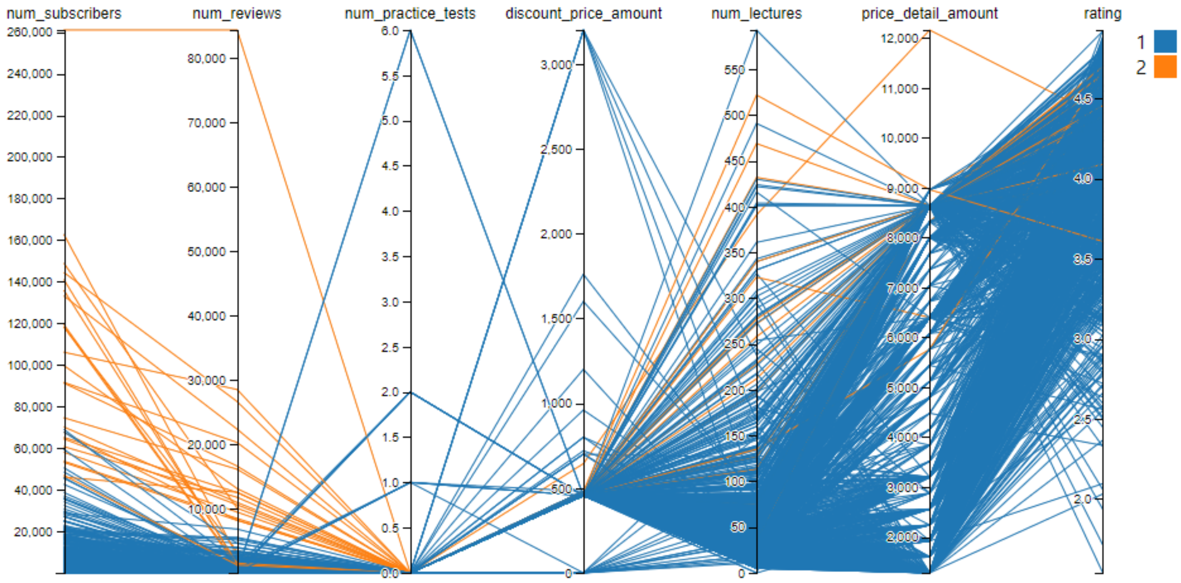
- On clicking on the variable points, user can select the ordering on PCP axes

Correlation MDS Scatterplot



8. Parallel Coordinate Plot
    - Helps in visualizing all the dimension attributes at a time
    - User can select an axis and place it wherever suitable for him
    - It also has brushing functionality to get good insights

Parallel Coordinate Plot



Technologies used:

- Flask is used as a backend server for data pre-processing, sampling, performing dimensionality reduction using PCA, MDS, k-means clustering
- D3.js along with HTML and CSS is used for plotting all the charts on frontend.
- Python version: 3.8.5, D3.js version 4

Code Execution:

- Open the project in environment having python and type ‘python app.py’ on terminal
- Go to Chrome browser at ‘http://127.0.0.1:5000/’

Video Link: https://youtu.be/nNnTAqyxXLU