CSE564 - VISUALIZATION - LAB ASSIGNMENT 2

(Full report)

Kenil Patel - 113262209

College Basketball Data Analysis

Dataset:

College basketball dataset taken from Kaggle:
https://www.kaggle.com/andrewsundberg/college-basketball-dataset

Objective:

- **Task 1**: **Basic dimension reduction and data visualization with PCA**
  - Computing EigenVectors and eigenvalues for the data using PCA and plotting the scree plot (1.1)
  - Adding interaction element to allow user to select the intrinsic (1.2) dimensionality index on the scree plot
  - Plotting the data into biplot (1.3)
- **Task 2**: **Visualization of data using scatter plot matrix**
  - Selecting the 4 PCA components which are less than the dimensionality index selected in task 1 and listing it in the table (2.1)
  - Constructing scatter plot matrix using the above 4 components (2.2)
  - Finding clusters using k-means and coloring them by color id (2.3)
- **Task 3**: **MDS (Multidimensional Scaling) plots**
  - Construct a MDS plot using the Euclidean distance and visualize it via a scatter plot (3.1)
  - Construct the MDS plot using 1 - |correlation| distance and visualize it via a scatter plot (3.2)
- **Task 4**: **Parallel coordinates plot (PCP)**
  - Visualize the data in a parallel coordinates plot (4.1)
  - Color the polylines by cluster ID (4.2)

Attributes:

- # of games played (ngames): Total number of games that the team played
- # of games won (wgames): Total number of games that the team won
- Adjusted Offensive Efficiency (off_eff) - It refers to the estimate of the offensive efficiency (points scored per 100 possessions) a team would have against the average Division 1 defense
- Adjusted Defensive Efficiency (def_eff) - It refers to the defensive efficiency (points allowed per 100 possessions) a team would have against the average Division 1 offense
- Power Rating (rating) - Chance of beating an average Division 1 team)
- Effective Goal Percentage Shot (egps)
- Effective Goal Percentage Allowed (egpa)
- Turnover Percentage Allowed (tpa)
- Turnover Percentage Committed (tpc)
- Offensive Rebound Rate (orr)
- Offensive Rebound Rate Allowed (ora)
- Free Throw Rate (ftr) - How often the given team shoots free throw
- Free Throw Rate Allowed (fta)
- Two Point Shooting Percentage (tps)
- Two Point Shooting Percentage Allowed (tpa)
- Three Point Shooting Percentage (thp)
- Three Point Shooting Percentage Allowed (thpa)
- Adjusted Tempo (at) - An estimate of the tempo (possessions per 40 minutes) a team would have against the team that wants to play at an average Division I tempo)
- Wins Above Bubble (wab) - The bubble refers to the cut off between making the NCAA March Madness Tournament and not making it

Implementation:

Backend:

- Data is cleaned using the jupyter notebook (handling missing values, string to int conversion, etc). The cleaned dataset is stored in the data folder of the code and is titled ‘basketball_data.csv’

- **Flask Server**: The cleaned data is loaded into the flask server, which does the following tasks:
    1. Filters out the numerical attributes from the pandas dataframe and create attribute to index mapping

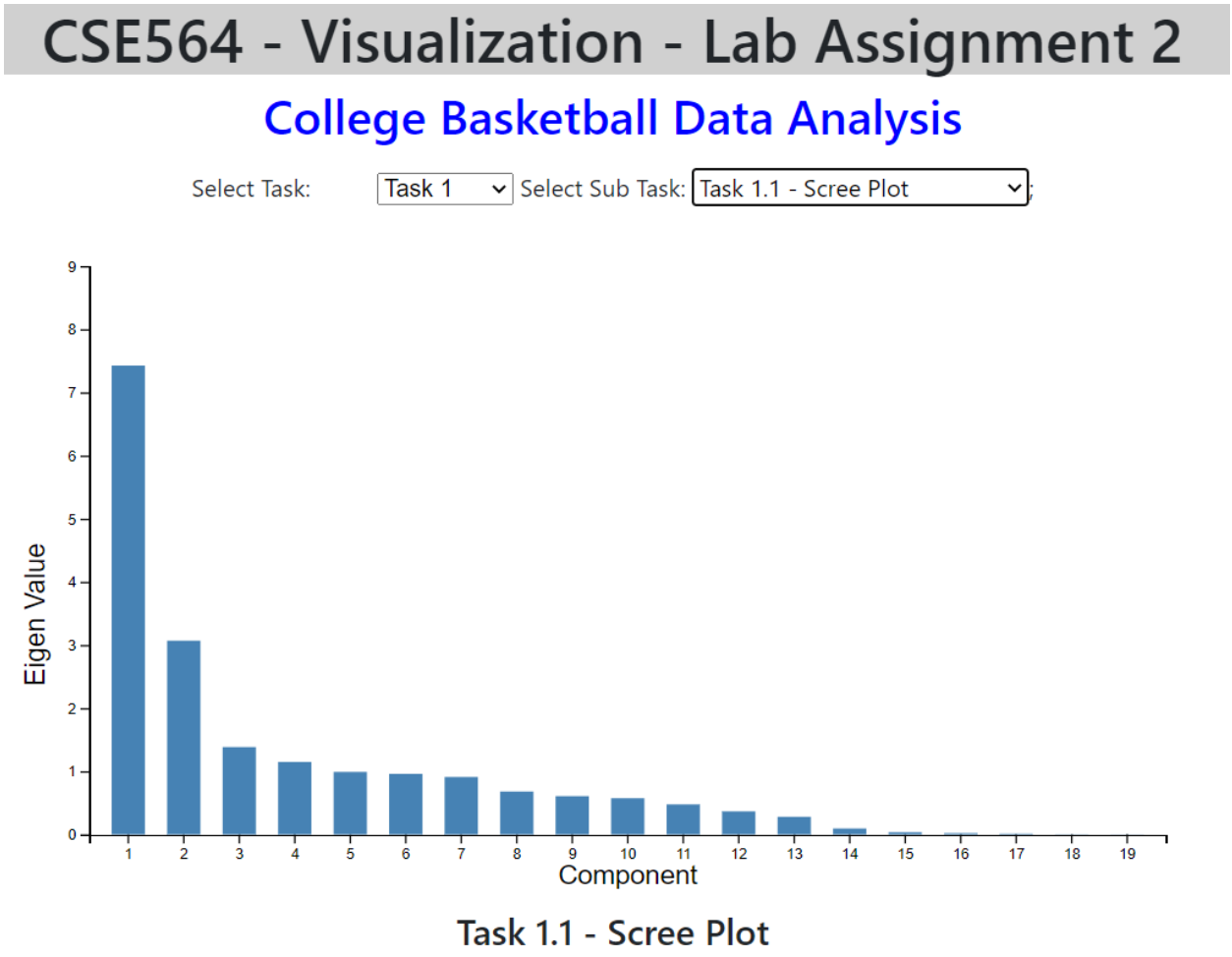
2. The pandas dataframe is then fed to the StandardScaler() function which rescales the data in such a way that the mean equals 0 and variance equals 1. (standardization).
3. Compute PCA of the standardized data (used sklearn to compute) which gives eigenvectors as the output. Each eigenvector has its corresponding eigenvalue.
4. Sort the eigenvalues and its corresponding eigenvector. The highest eigenvalue corresponds to the highest variance and the smallest eigenvalue indicates least variance.
5. For task 2, factor loadings are calculated for plotting the tabular data and scatter plot matrix
6. kMeans clustering is implemented using the python's inbuilt kMeans function which returns clusters of identical properties.
7. Since kmeans clustering is computation intensive for a large number of data points, data for MDS plot and PCP plot has been precomputed as soon as the window loads to save time.

- **Visualization:**
    1. The js file receives data from the flask server on the go and it plots the scree plot, biplot and the scatter plot matrix based on the user selection.
    2. The user selects the intrinsic dimensionality index value on the scree plot and that value is passed to the flask app, which then uses it to populate the top 4 attributes in the table.
    3. PCA Biplot - The biplot is the combination of the PCA score plot and the loadings plot. The score plot graphs the score of PCA1 vs PCA2. The loading plot on the other hand graphs the coefficient of each variable for the first component vs the coefficient of the second component. Loadings can range from -1 to 1.
    4. The scatter plot matrix has 4 x 4 subplots where the text in the subplot represents the attribute on y axis and the text value in the row represents the attribute on x axis.
    5. The Parallel Coordinate Plot represents all the data points and their values are represented on the vertical axis. Each datapoint has a unique color depending on the cluster assigned to it.

Observations:

1) In the biplot, the majority of the data seems to follow normal distribution as evident from the fact that a huge collection of points is randomly distributed around zero. The points away from zero are the outliers.
2) Loadings close to -1 or 1 indicates that that variable strongly influences the principal component. Loadings which are closer to 0 indicates that that variable has a very weak influence on the principal component.

Screenshots:

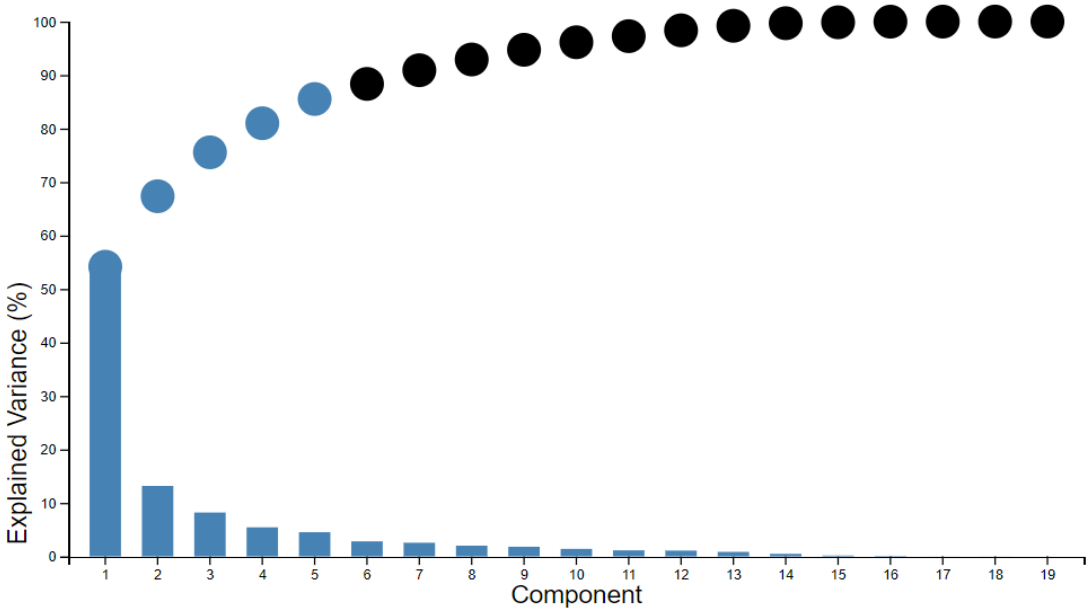


CSE564 - Visualization - Lab Assignment 2

College Basketball Data Analysis

Select Task: Task 1 Select Sub Task: Task 1.2 - Explained Variance

You have selected di = 4



Task 1.1 - Scree Plot

CSE564 - Visualization - Lab Assignment 2

College Basketball Data Analysis

Select Task: Task 2 Select Sub Task: Task 2.1 - Table Data

4 attributes with highest PCA loadings

Selected Attribute	Sum of Squared Loadings
off_eff	0.9725362393944972
def_eff	0.8483563142317929
fta	0.5010870153832436
ftr	0.40147304574495996

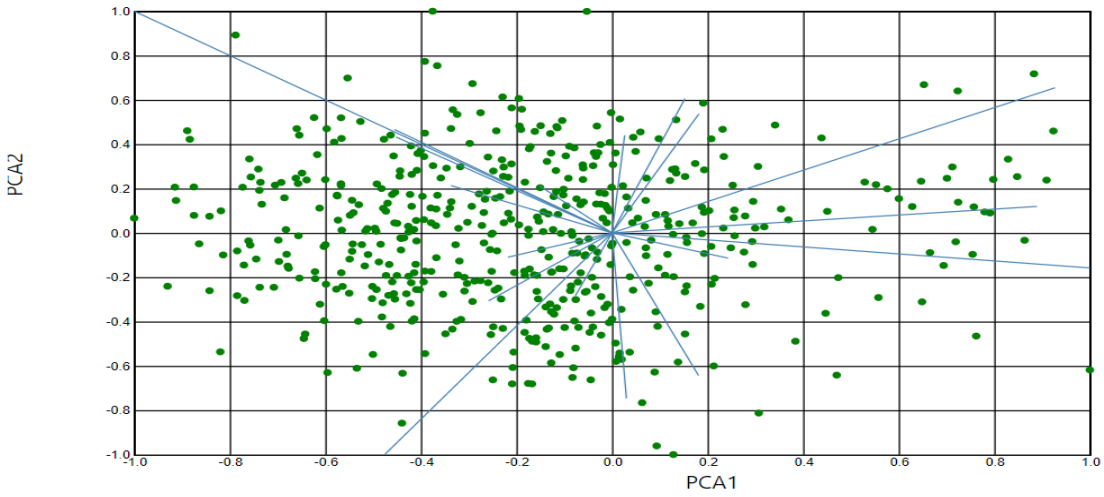
Task2.1 - Table Data(4 attributes with highest PCA loadings)

CSE564 - Visualization - Lab Assignment 2

College Basketball Data Analysis

Select Task: Task 1 Select Sub Task: Task 1.3 - PCA Biplot

You have selected di = 4



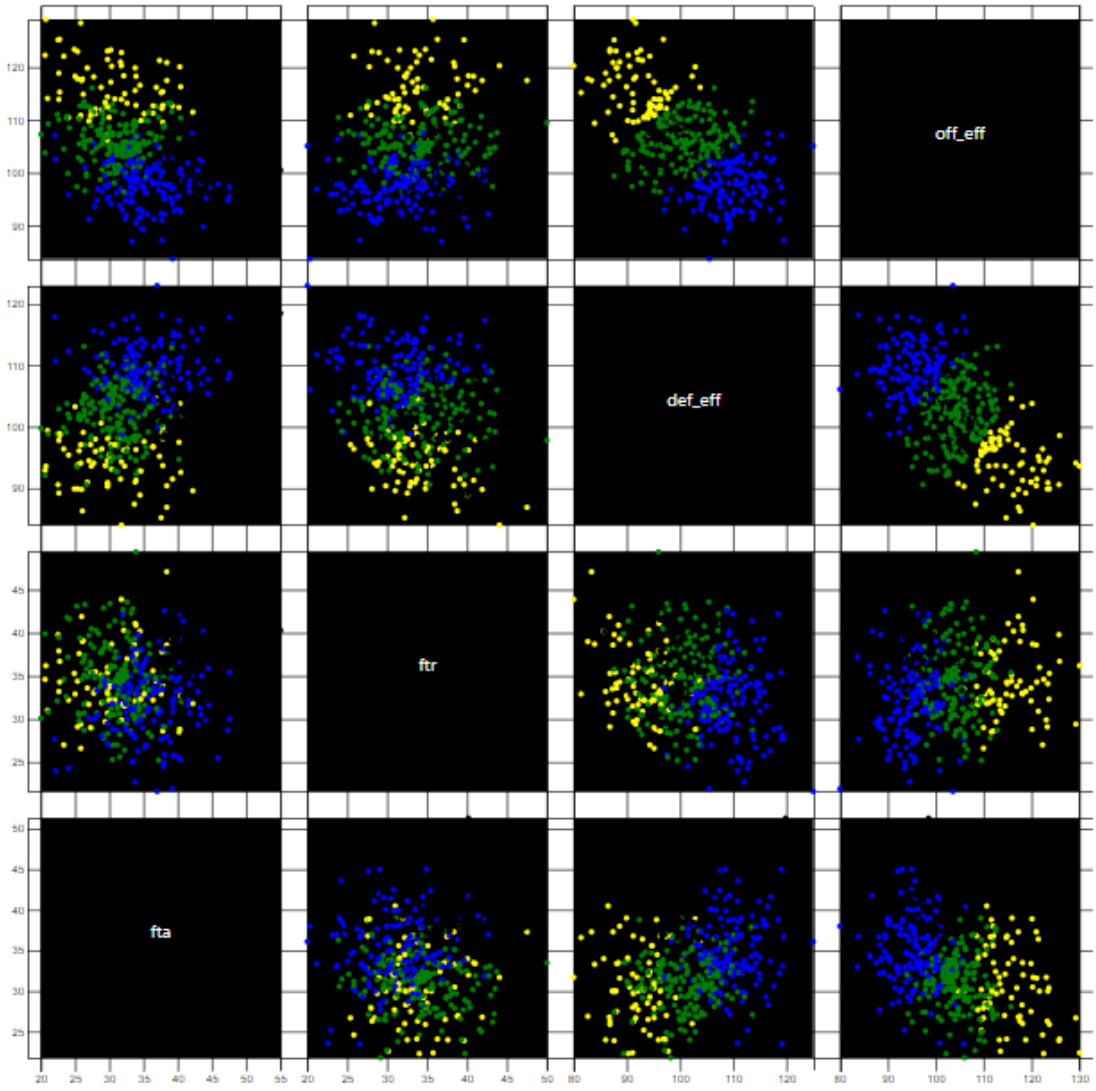
Task1.3 - PCA biplot = PCA score plot + Loading plot

CSE564 - Visualization - Lab Assignment 2

College Basketball Data Analysis

Select Task: Task 2 Select Sub Task: Task 2.2 - Scatter Plot Matrix

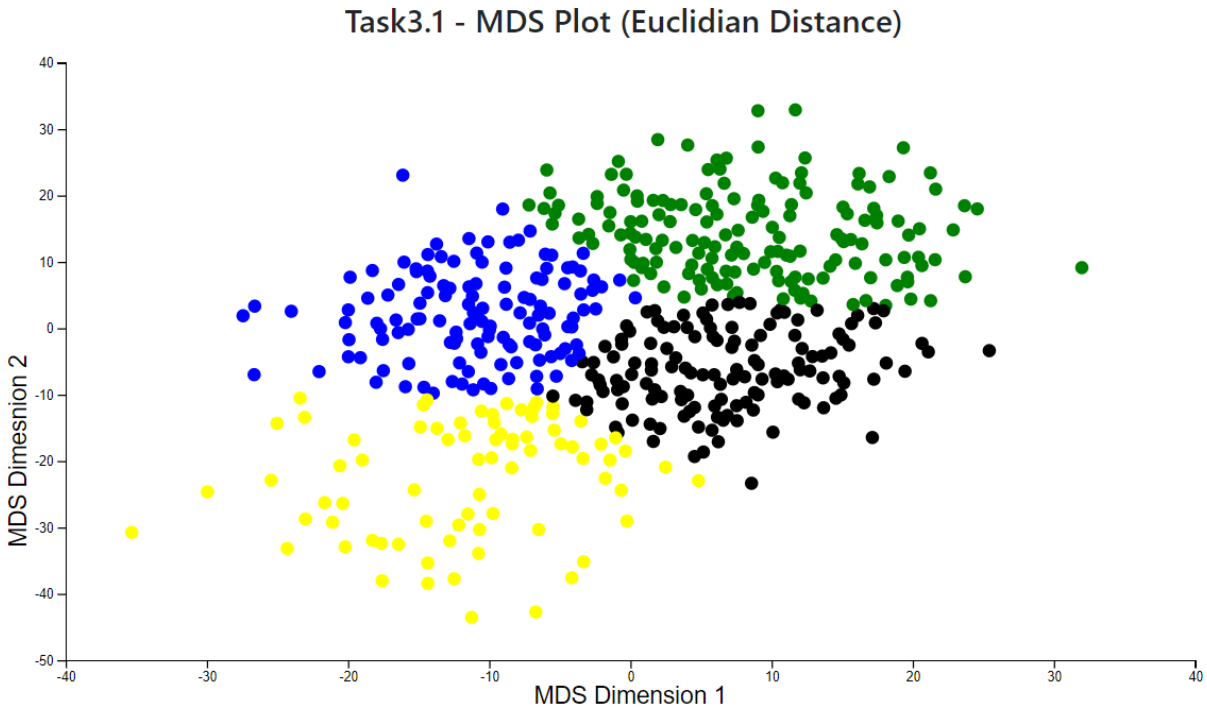
Task 2.2 - Scatter Plot Matrix



CSE564 - Visualization - Lab Assignment 2

College Basketball Data Analysis

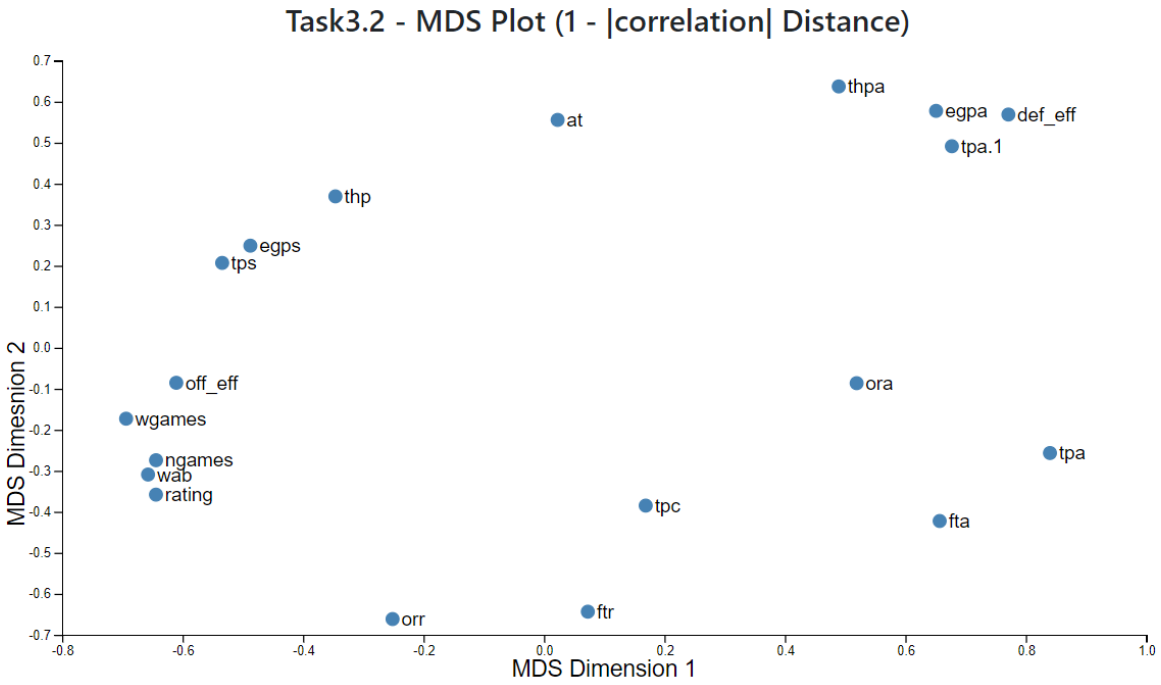
Select Task: Task 3 Select Sub Task: Task 3.1 - MDS plot (Euclidian distance)



CSE564 - Visualization - Lab Assignment 2

College Basketball Data Analysis

Select Task: Task 3 Select Sub Task: Task 3.2/3 - MDS plot (1-|correlation| distance)



CSE564 - Visualization - Lab Assignment 2

College Basketball Data Analysis

Select Task: Task 4 Select Sub Task: Task 4.1 - PCP plot