

# CSE564 - VISUALIZATION - LAB ASSIGNMENT 2

## (Task 1 & 2)

Kenil Patel - 113262209

### College Basketball Data Analysis

#### 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)

#### Attributes:

- # of games played (G): Total number of games that the team played
- # of games won (W): Total number of games that the team won
- Adjusted Offensive Efficiency (ADJOE) - 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 (ADJDE) - It refers to the defensive efficiency (points allowed per 100 possessions) a team would have against the average Division 1 offense
- Power Rating (BARTHAG) - Chance of beating an average Division 1 team)
- Effective Goal Percentage Shot (EFG\_O)
- Effective Goal Percentage Allowed (EFG\_D)
- Turnover Percentage Allowed (TOR)
- Turnover Percentage Committed (TORD)
- Offensive Rebound Rate (ORB)
- Offensive Rebound Rate Allowed (DRB)

- Free Throw Rate (FTR) - How often the given team shoots free throw
- Free Throw Rate Allowed (FTRD)
- Two Point Shooting Percentage (2P\_O)
- Two Point Shooting Percentage Allowed (2P\_D)
- Three Point Shooting Percentage (3P\_O)
- Three Point Shooting Percentage Allowed (3P\_D)
- Adjusted Tempo (ADJ\_T) - 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:

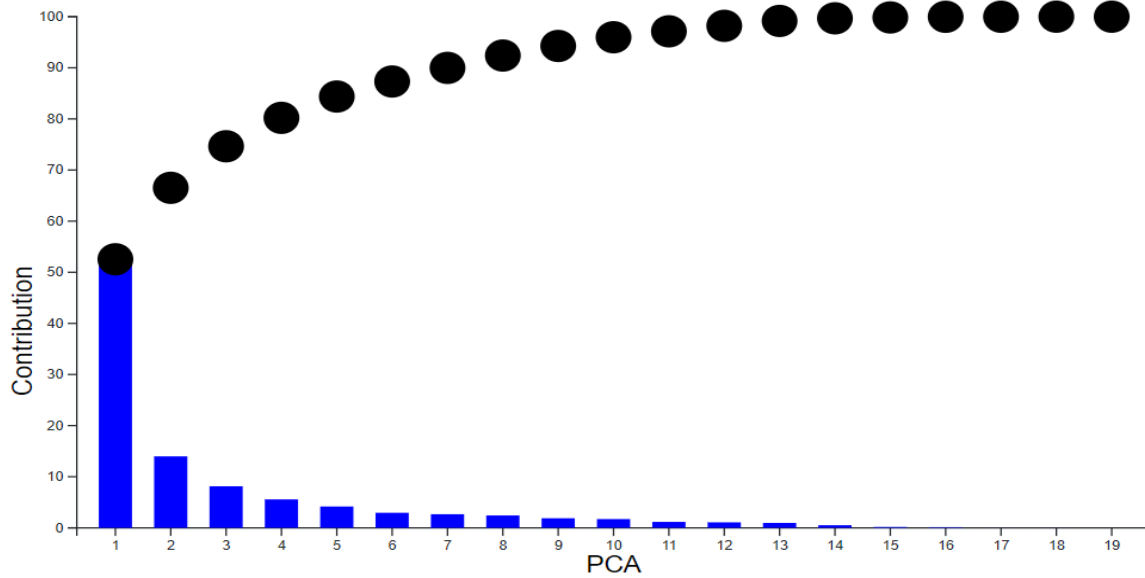
- Data is cleaned using the jupyter notebook (handling missing values, string to int conversion, etc)
- **Flask Server:** The cleaned data is loaded into the flask server, which does the following tasks:
  1. Filters out the numerical attributes from the pandas dataframe and create attribute to index 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.
- **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. 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.

## Screenshots:

SBU ID: 113262209

# CSE564 - Visualization - Lab Assignment 2

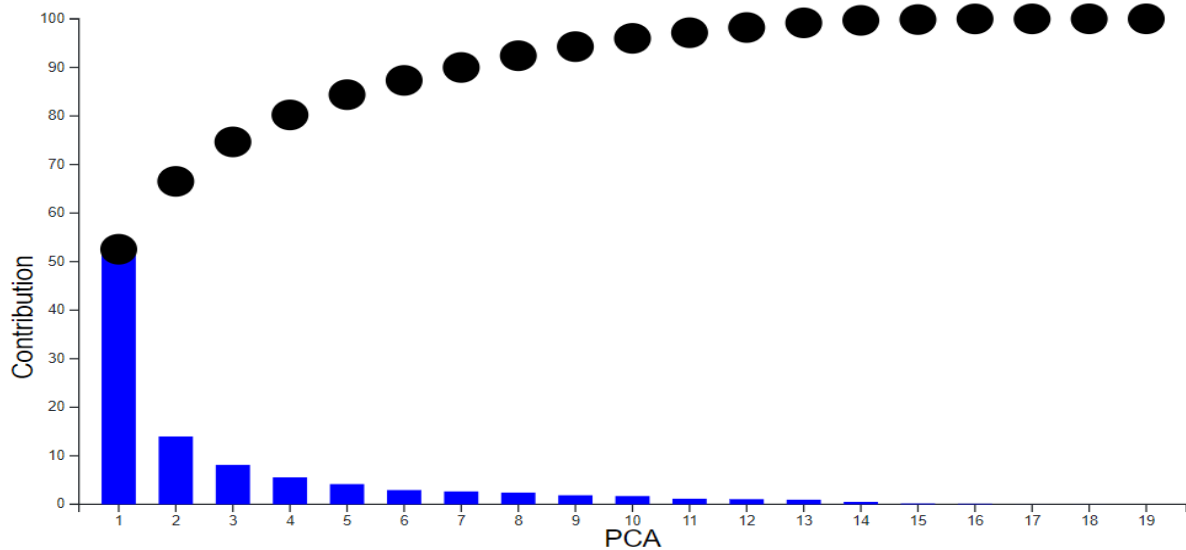
Select Task: Task 1 Select Sub Task: Scree Plot of Eigen Values



Task 1.1 - Visualizing EigenVectors as Scree Plot

## CSE564 - Visualization - Lab Assignment 2

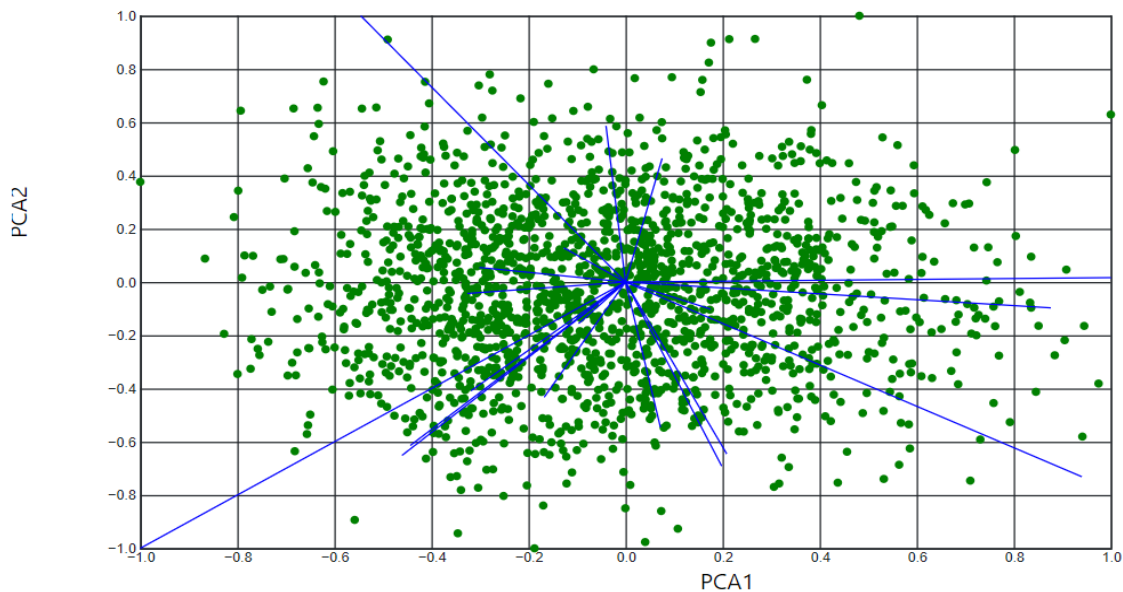
Select Task:  Select Sub Task:   
Selected Intrinsic\_d = 4



Task 1.2 - User selects intrinsic dimensionality index

## CSE564 - Visualization - Lab Assignment 2

Select Task:  Select Sub Task:



Task 1.3 - PCA based biplot

SBU ID: 113262209

# CSE564 - Visualization - Lab Assignment 2

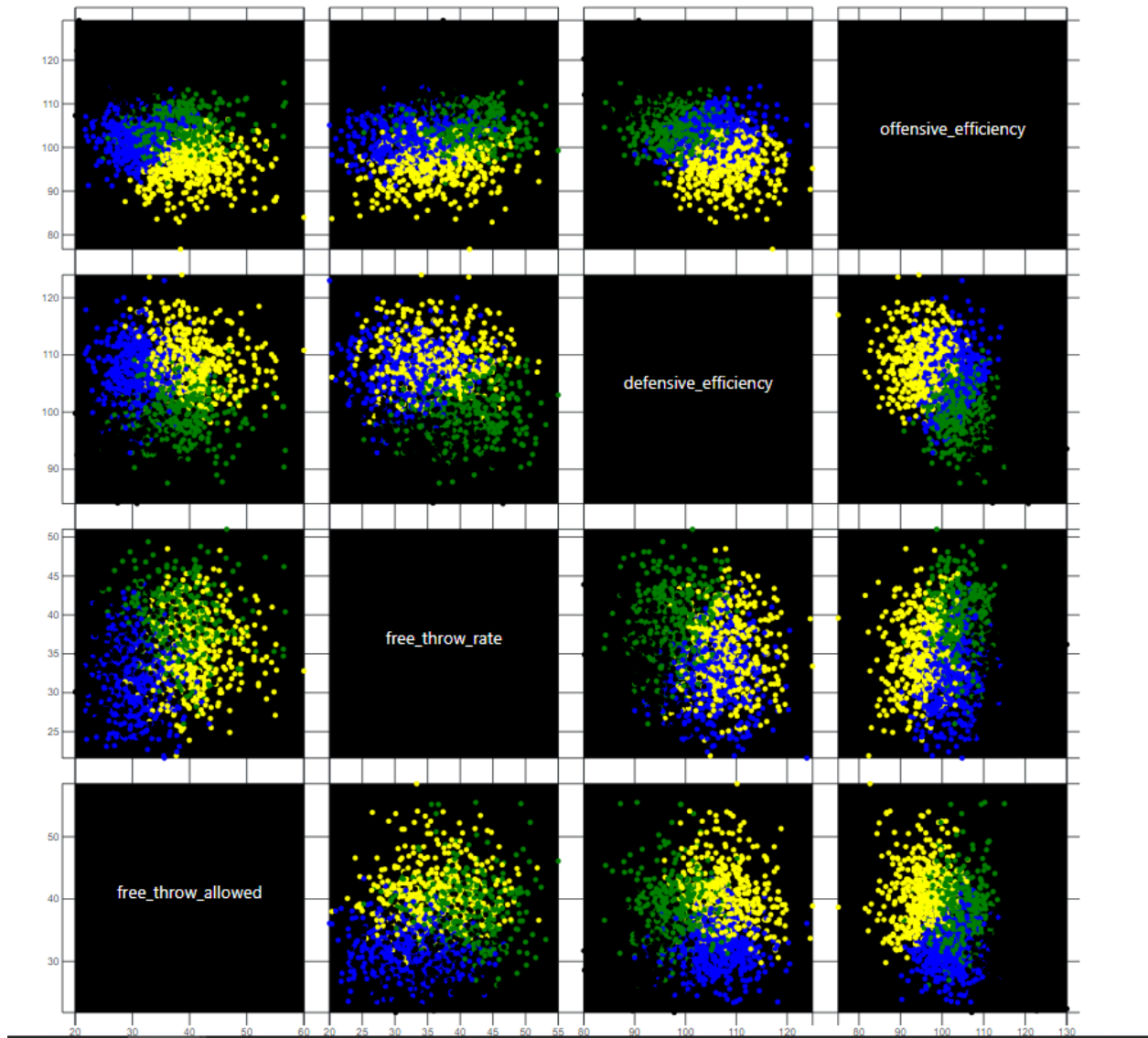
Select Task: Task 2 ▼ Select Sub Task: Table Data ▼

Selected Attribute Sum of Squared Loadings	
defensive_efficiency	0.8282544907717355
free_throw_allowed	0.4722677050129794
free_throw_rate	0.456390327412224
offensive_efficiency	0.9375785085261087

Task 2.1 - PCA attributes with highest PCA loadings

## CSE564 - Visualization - Lab Assignment 2

Select Task:  Select Sub Task:



Task 2.2 & 2.3 - PCA attributes with highest PCA loadings, k-means clusters of data