# CSE 564 Visualization Lab 2 Report

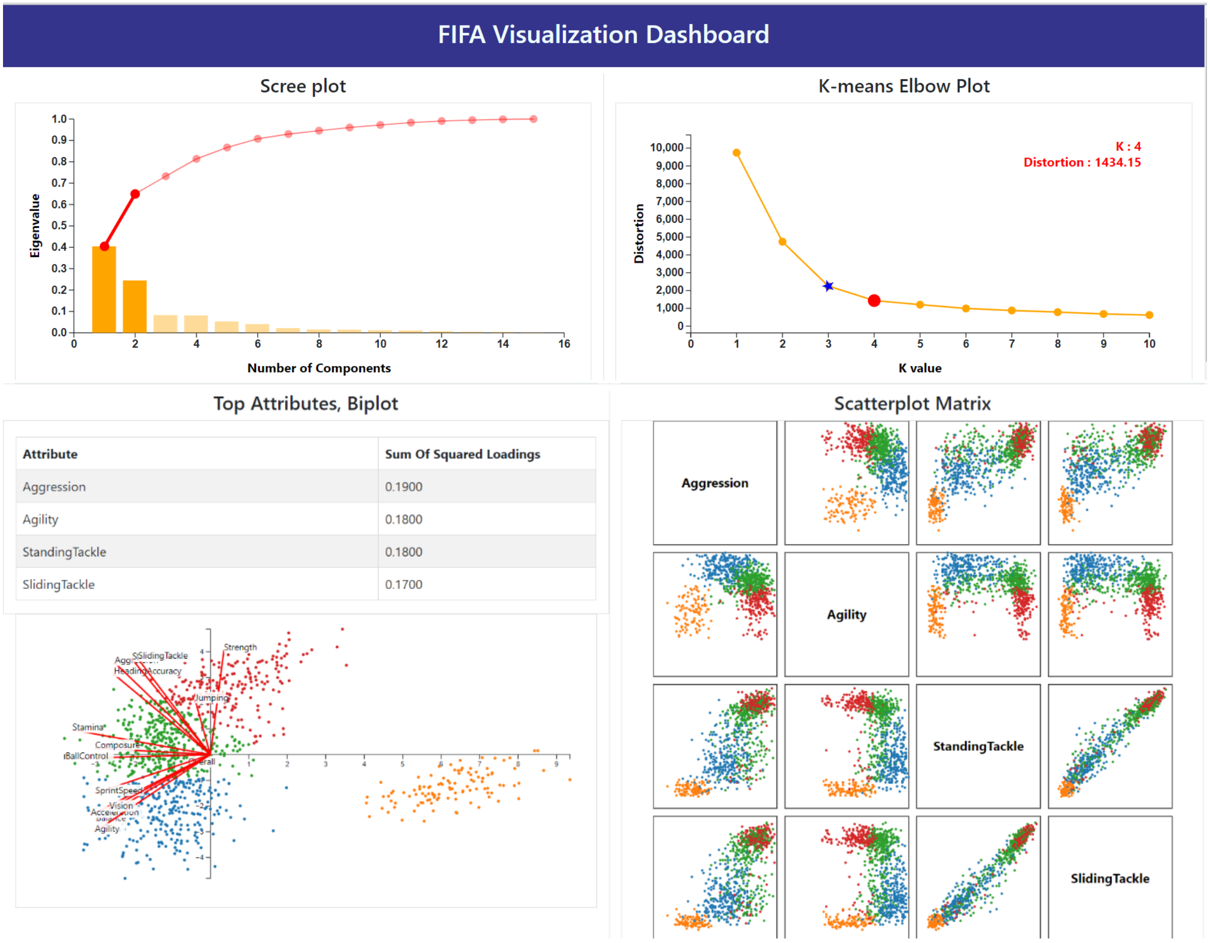
**DATASET**: FIFA complete player dataset (https://www.kaggle.com/karangadiya/fifa19) This dataset contains detailed attributes for every player registered in the FIFA 19 database. It has 18.2k rows and 89 attributes. A large number of the attributes are numerical. I have chosen 15 numerical attributes for the purpose of this assignment.

The FIFA 19 Complete Player Dataset is fascinating due to its comprehensive coverage of over 18.2k players and 89 attributes, including technical and physical skills. This depth enables detailed analyses of player diversity, playing styles, and factors contributing to virtual soccer success. Moreover, its integration of real-world player statistics adds authenticity, appealing to both gaming enthusiasts and sports analysts. Additionally, the dataset's numerical attributes facilitate comparative analyses and directly impact gameplay mechanics, offering insights for gamers optimizing team compositions and developers refining gameplay balance and realism. In summary, this dataset provides valuable insights into virtual soccer dynamics, blending gaming, sports, and data analysis.

They are as follows:

- Overall score
- Balance
- Stamina
- Strength
- Heading Accuracy
- Ball Control
- Acceleration
- Sprint Speed
- Agility
- Aggression
- Jumping
- Vision
- Composure
- Standing Tackle
- Sliding Tackle

## DASHBOARD INTERFACE DESIGN:



FIFA Visualization Dashboard

Scree plot

K-means Elbow Plot

Top Attributes, Biplot

| Attribute | Sum Of Squared Loadings |
|---|---|
| Aggression | 0.1900 |
| Agility | 0.1800 |
| StandingTackle | 0.1800 |
| SlidingTackle | 0.1700 |

Scatterplot Matrix

**Scree Plot:** The Scree Plot is interactive and allows users to explore the intrinsic dimensionality of the dataset by selecting the number of components (1-15). This plot displays the variance explained by each principal component, aiding users in determining the optimal number of components to retain for analysis. As users adjust the dimensionality index, other charts such as the Attribute Table, BiPlot, K-means Elbow Plot, and Scatter Plot Matrix are dynamically updated to reflect changes.

**Attribute Table:** The Attribute Table showcases the top 4 attributes with the highest loadings and their sum of squared loadings, offering insights into the most influential features in the dataset.

**BiPlot:** The BiPlot is a scatterplot of PC1 vs PC2, augmented with loading vectors representing the contributions of original variables to the principal components. This plot visually illustrates the relationships between variables and principal components, enabling users to discern patterns and correlations within the dataset. By examining the direction and magnitude of loading vectors, users can gain insights into the underlying structure and dimensionality of the data.

**K-means Elbow Plot:** The K-means Elbow Plot illustrates the distortion (inertia) for k values ranging from 1 to 10, allowing users to identify the optimal number of clusters for their data. By plotting distortion against the number of clusters, users can observe the "elbow point," indicating the most significant decrease in distortion. This point, denoted by a blue star on the curve, suggests an appropriate number of clusters for further analysis. The K-means Elbow Plot is interactive and guides users in selecting the optimal k value, influencing subsequent clustering and visualization in other plots.

**Scatter Plot Matrix:** The Scatter Plot Matrix presents pairwise scatterplots of the top 4 attributes based on the PCA with the selected dimensionality index. While this graph is not interactive, it offers a visual representation of the relationships between variables, highlighting potential clusters or patterns within the data. Additionally, the colors of data points in the Scatter Plot Matrix change based on the k value selected in the K-means Elbow Plot, providing users with additional insights into clustering structures and groupings within the dataset.

## FLASK APP:

The backend of this web application is built using Python and leverages several key libraries to handle data processing, computation, and routing functionalities. The primary libraries used include pandas for data manipulation, numpy for numerical computations, scikit-learn for PCA and K-means clustering algorithms, and Flask for web development.

Upon initializing the Flask application, routes are defined to handle different types of requests from the frontend. For example, the '/' route renders the main index.html template, while '/pca_data', '/elbow_plot_data', and '/pca_idi_data' routes serve specific data for PCA, K-means Elbow Plot, and PCA with Intrinsic Dimensionality Index (IDI), respectively.

In the '/pca_data' route, the backend performs Principal Component Analysis (PCA) on the standardized dataset to calculate explained variance ratios and prepare data for the Scree Plot. The '/elbow_plot_data' route computes the distortion values for different numbers of clusters (k) to generate data for the K-means Elbow Plot. Meanwhile, the '/pca_idi_data' route handles requests related to PCA with IDI, extracting top attributes, performing K-means clustering, and preparing data for the BiPlot and Scatter Plot Matrix.

## OBSERVATIONS:

1. In the Scree plot, the 'elbow' seems to occur at the third component, which suggests that the first two components explain the most significant portion of the variance in the data.

2. The fact that the first component alone accounts for 40% of the variance indicates that there is a strong dominant pattern in the dataset. The second component adds another 25% (totaling 65%), which is a significant amount of information. From the third component onwards, each additional component contributes a decreasing amount of additional explained variance (8%, 7%, 6%, etc.).

3. The blue star represents the calculated elbow point, which is where the rate of decrease sharply changes, indicating the appropriate number of clusters for the data. In this plot, the elbow point is at k = 3 which means four clusters are suggested by this method.

4. The distortion decreases as the number of clusters increases because the data points are closer to the centers of their respective clusters. However, after a certain point, the decrease in distortion is marginal, which suggests that adding more clusters does not provide a substantially better fit.

5. Jumping (0.9400): This attribute has the highest sum of squared loadings, which means it contributes significantly to the variance in the dataset. It's likely to be a dominant feature in the PCA, strongly influencing the direction and magnitude of the principal components. Overall (0.9000): The 'Overall' attribute is also a major contributor to the variance in the data, just slightly less than 'Jumping'. It is an important attribute for the principal components.

6. Jumping and Overall score: There seems to be a positive relationship between Jumping and Overall; as the jumping ability increases, the overall attribute tends to increase as well. This suggests that individuals who are good at jumping may also tend to have a higher overall score.

7. There isn't a clear linear pattern between Jumping and Vision, suggesting a weaker correlation between these attributes. Similarly the relationship between Vision and Strength does not show a clear linear trend, suggesting that these two attributes are less directly correlated compared to others.

8. Sliding Tackle and Strength: The scatterplot indicates some degree of positive correlation, with stronger players also tending to have better sliding tackle ability. However, there is a wide spread in the data points, especially among players with mid-range strength, which suggests that while there is a tendency for stronger players to perform better sliding tackles, it's not a definitive rule.

9. Attributes and Components: Attributes such as "Overall", "Composure", "Ball Control", "Sprint Speed", "Vision", "Agility", and "Acceleration" are all clustered together and point in a similar direction, suggesting they are positively correlated and contribute similarly to the principal components. On the other hand, "Strength", "Jumping", "Heading Accuracy", "Sliding Tackle", and "Aggression"point in a different direction, suggesting they contribute differently to the data's variance and may represent different aspects of the data.

10. The first principal component (horizontal axis) may represent a kind of underlying factor related to technical skills (like ball control, vision, and speed), while the second principal component (vertical axis) might represent physicality (like strength and jumping).