CSE 564 Visualization Lab 2 Report

Dataset chosen: FIFA 19 complete player dataset (https://www.kaggle.com/karangadiya/fifa19)

This dataset contains detailed attributes for every player registered in the latest edition of FIFA 19 database. It has 18.2k rows and 89 attributes. A large number of the attributes are numerical. I have chosen 19 numerical attributes for the purpose of this assignment. They are as follows:

- Overall score
- Balance
- Stamina
- Strength
- Heading Accuracy
- Short Passing
- Long Passing
- Dribbling
- Ball Control
- Acceleration
- Sprint Speed
- Agility
- Shot Power
- Aggression
- Jumping
- Vision
- Composure
- Standing Tackle
- Sliding Tackle

Code structure:

My code consists of a Python backend (app.py) made in Flask, which does all the required computations like PCA, MDS, etc. The frontend consists of data visualizations made in D3.js. It is made of HTML, CSS and JavaScript files.

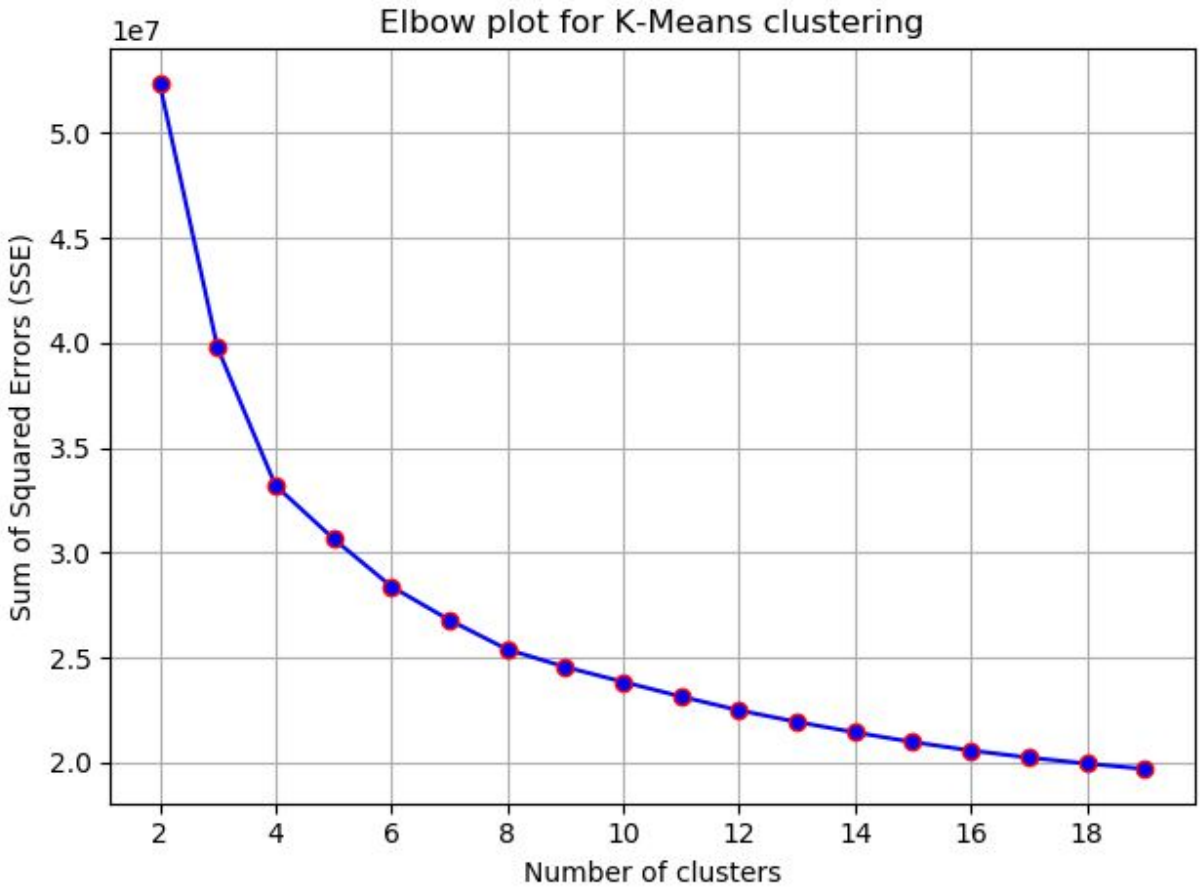
Task 1: Data clustering and decimation

Random sampling performed using random.sample in Python.

```
random_indices = random.sample(range(len(data_arr)), sample_size)
random_sampling_results = data_arr[random_indices]
```

Stratified sampling performed by first performing k-means clustering, and then performing random sampling on each individual cluster. The sample size of each individual cluster is set proportional to the size of the cluster. Number of clusters fixed by plotting elbow plot.

K-means performed using sklearn.cluster.KMeans.



Task 2: Dimension reduction

PCA is computed using sklearn.decomposition.PCA.

```
pca = PCA()
pca.fit(data)
pca_results = pca.explained_variance_ratio_
loadings = np.sum(np.square(pca.components_), axis=0)
indices_of_top_3_attributes = loadings.argsort()[-3:][::-1]
top_two_components = pca.components_[:2]
```

The three attributes with the highest PCA loadings are ‘Overall', 'Short Passing' and 'Sprint Speed'.

Observations:
The curve elbows at n_components=3. We see that out of 19 components, 3 components are enough to explain more than 75% variance in the data and 7 components are enough to explain more than 90% of the variance.

Task 3: Visualization
MDS is calculated using sklearn.manifold.MDS.
Dissimilarity matrix is calculated using scipy.spatial.distance.cdist.