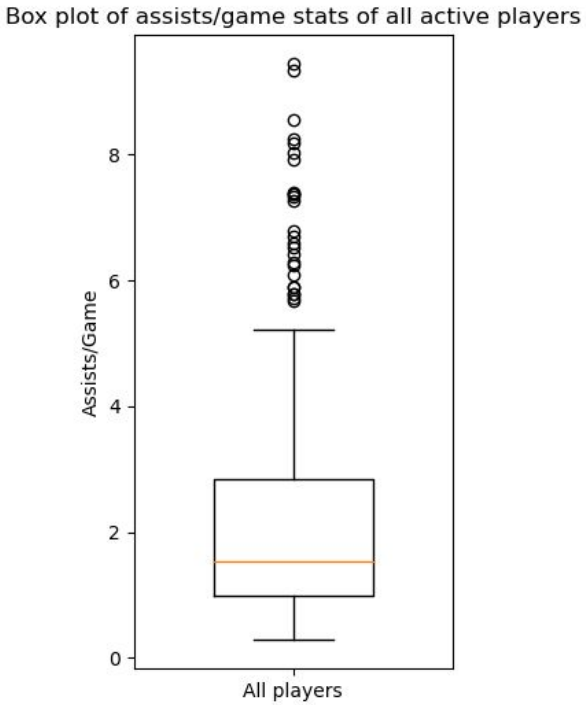
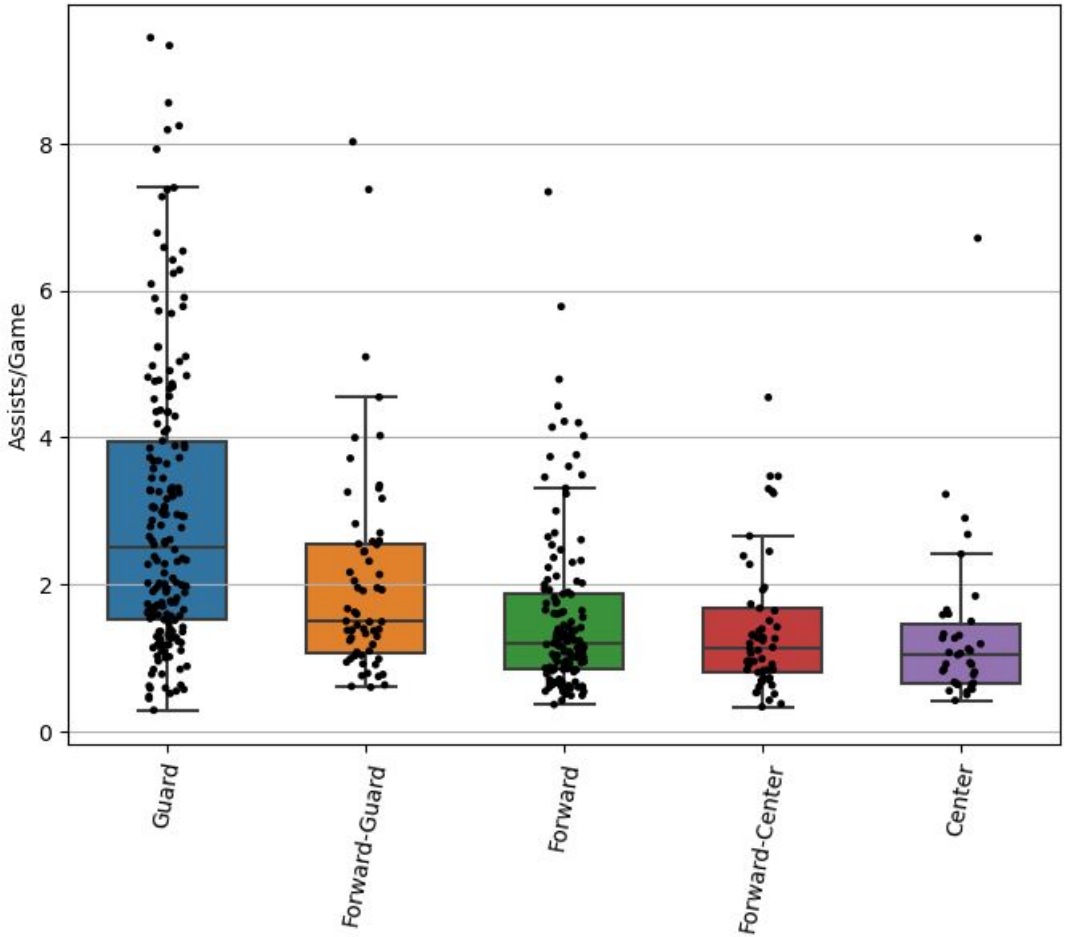
# NBA Data Analysis

Éliás Gergely

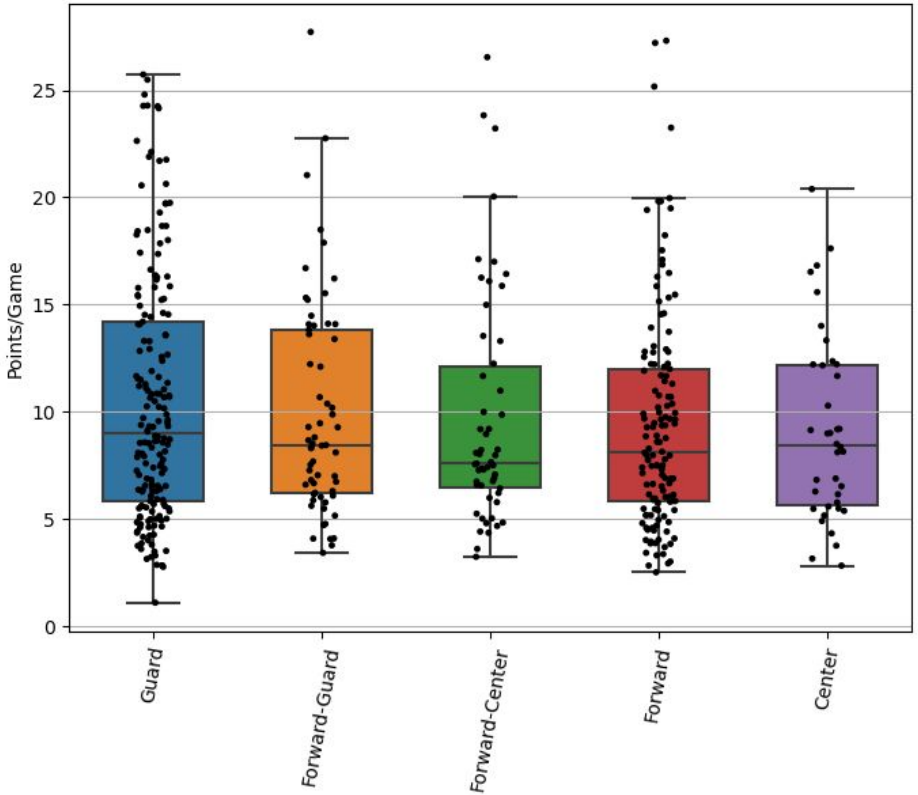
# Box plots of A/G

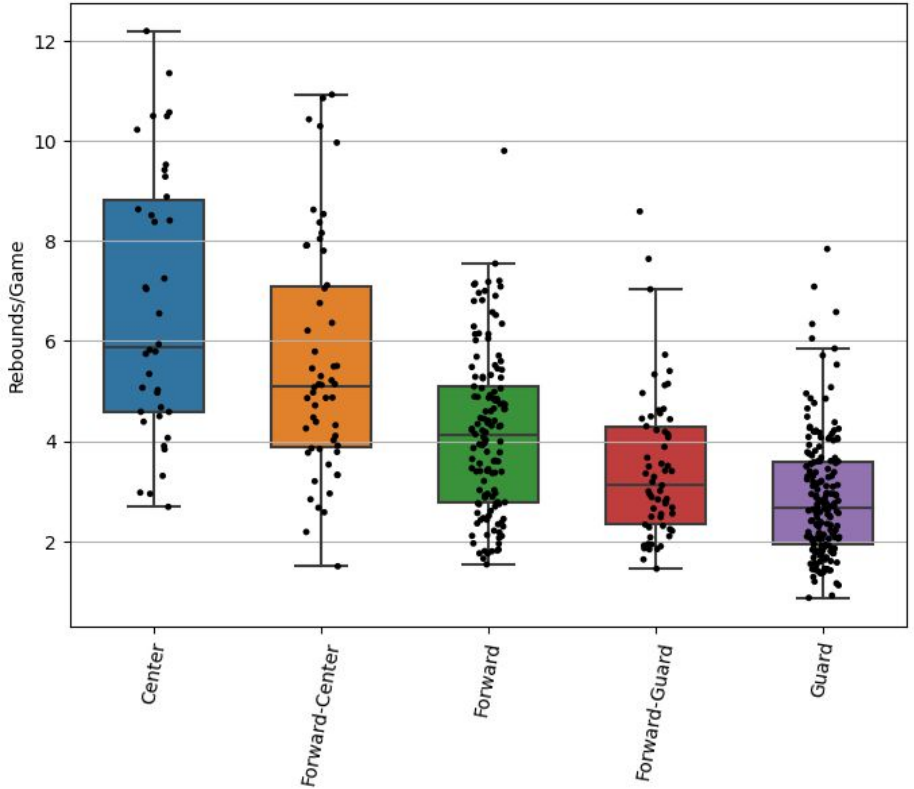


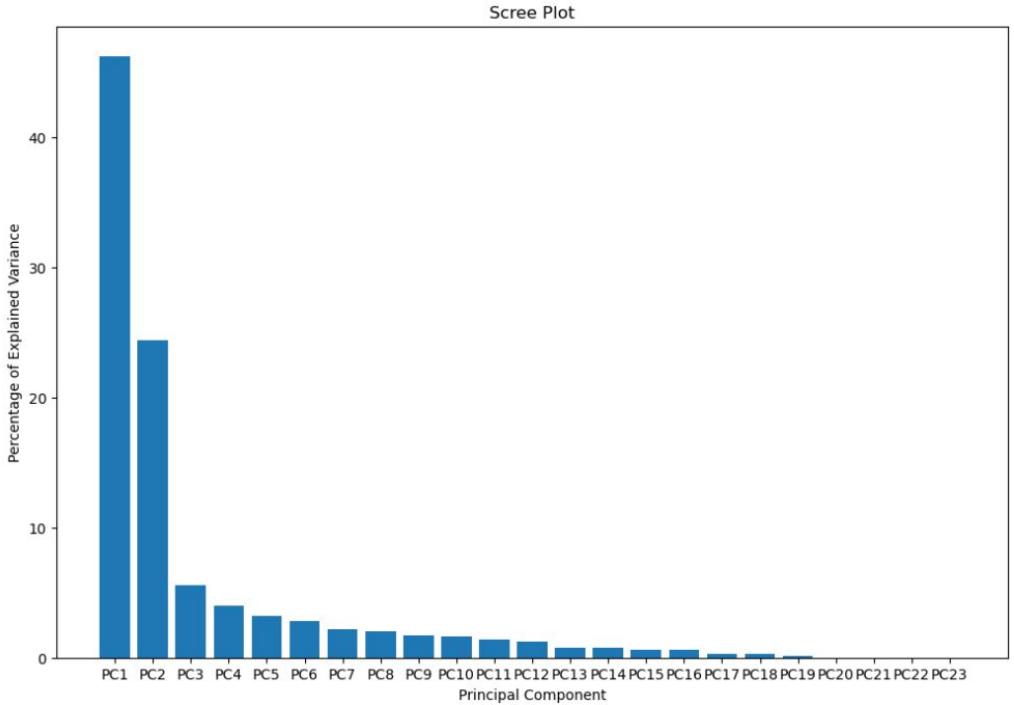

Guards perform best in terms of assists/game, but there are some outliers in other positions as well.

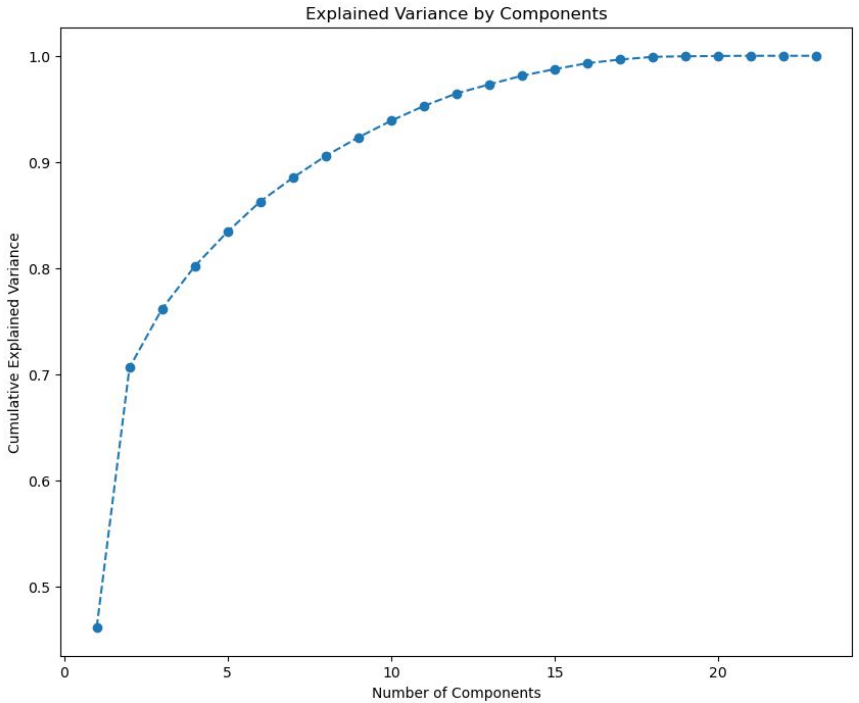
# Box plots of P/G and R/G per positions



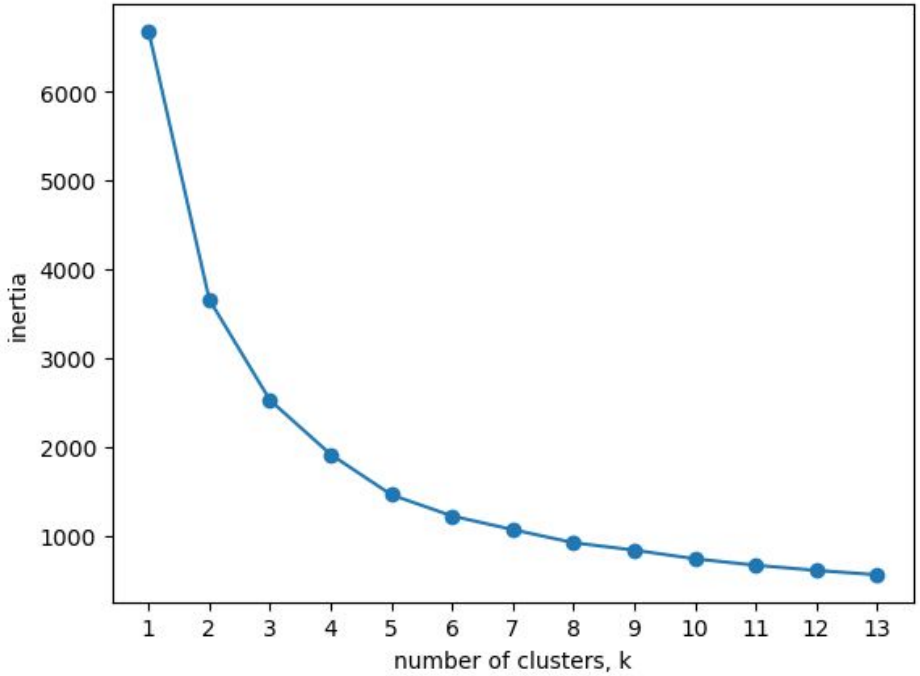
# PCA K-Means Clustering of NBA players



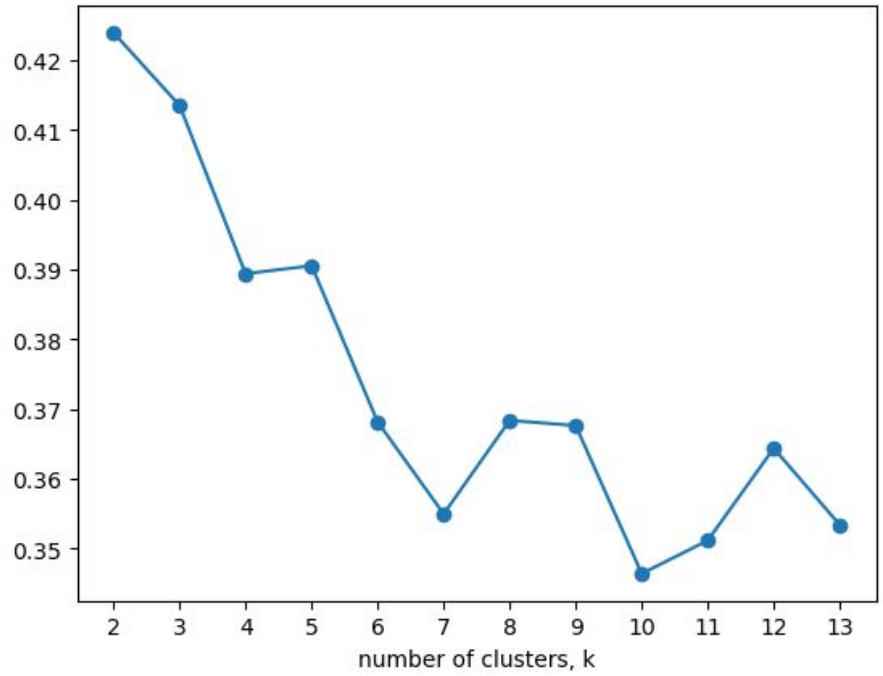




Analysis of variance explained by each principal components. PC1 and PC2 are the two components that capture the most variance in the dataset, together accounting for about 75%. That is why i chose two principal components for K-Means clustering.

# Elbow plot

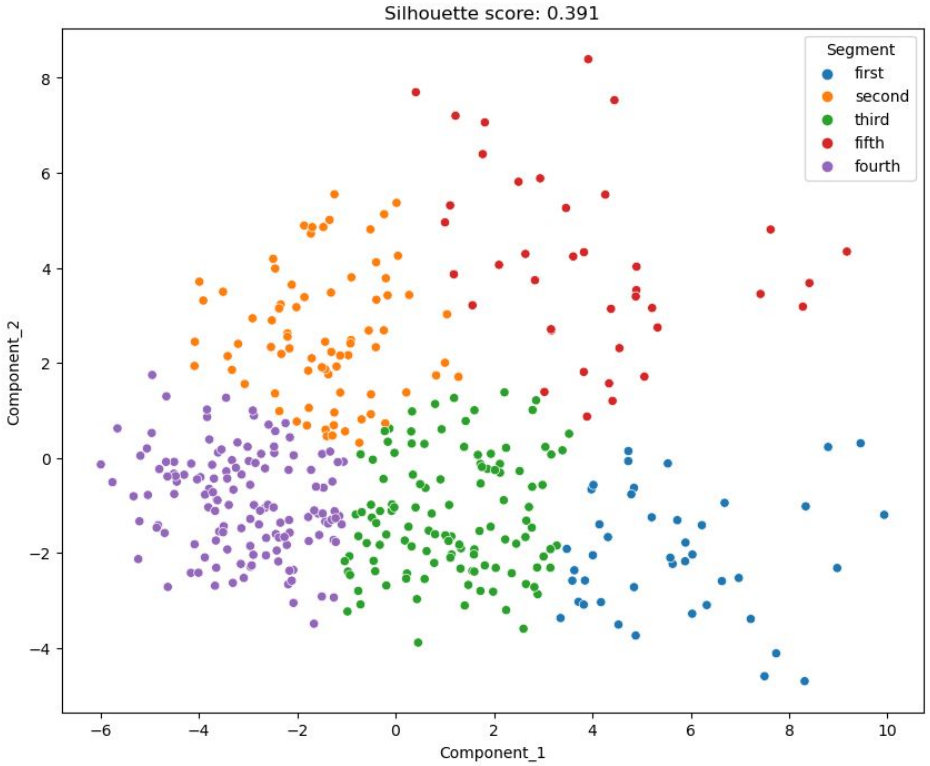



# Silhouette score plot

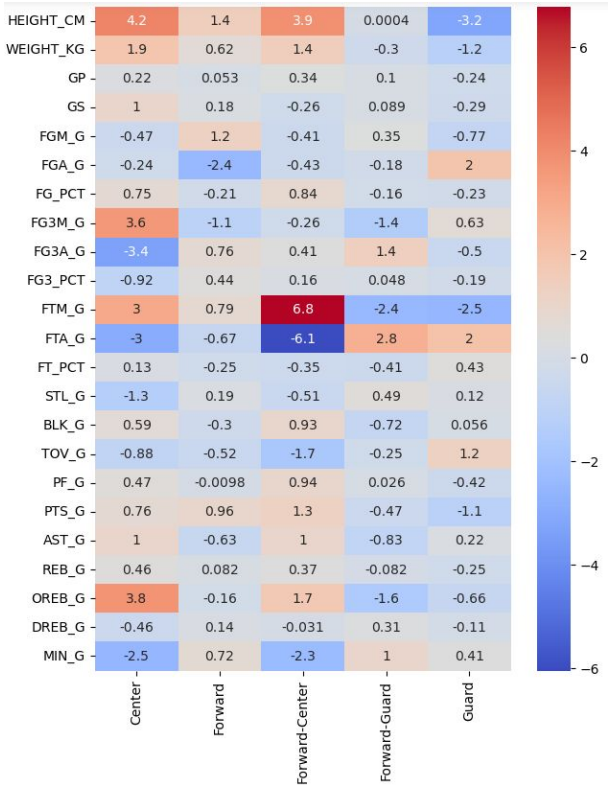



Evaluating the quality of the clustering using inertia and silhouette score.

# Scatterplot of the first to principal components, with the data points colored by their segment



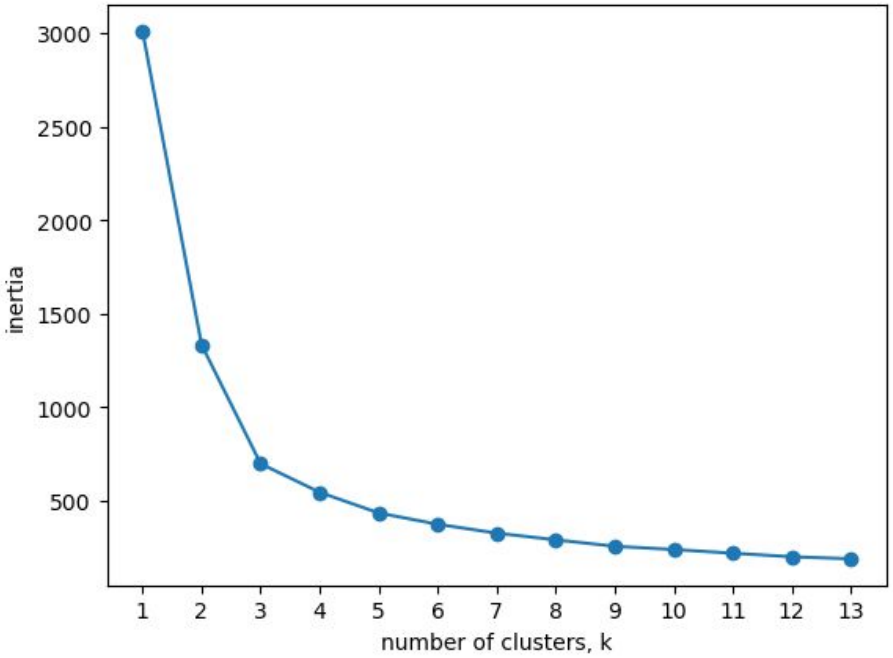
# LDA K-Means Clustering of NBA players



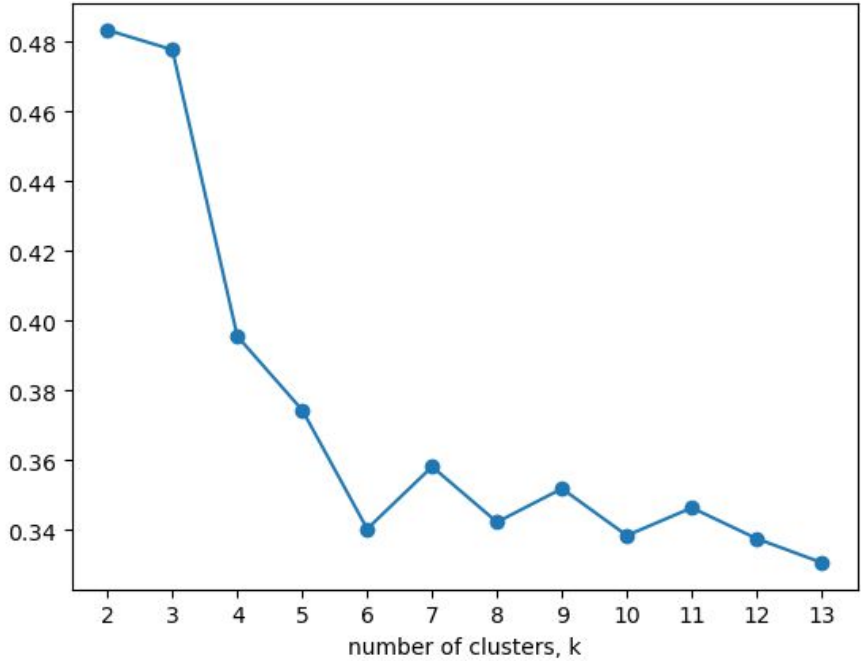


Visualization of the coefficients of the LDA model for each feature and position using a heatmap and visualization of the sum of the absolute values of the coefficients for each position.

# Elbow plot

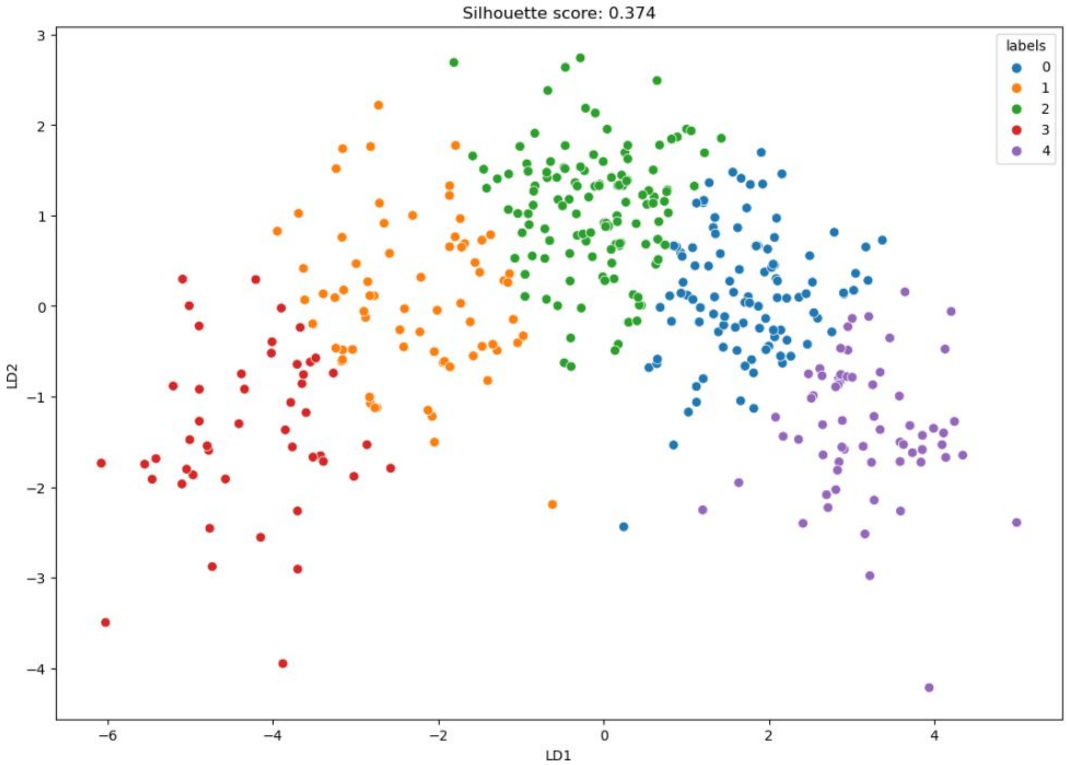



# Silhouette score plot




Evaluating the quality of the clustering using inertia and silhouette score.

# Scatterplot of the LDA-transformed data, with the data points colored by their segment



# Analyzing the accuracy of the clustering by comparing the assigned cluster labels with the actual positions

```
true = grouped5[grouped5['Check similarity'] == True]

accuracy = true['Position_based_on_clustering'].count() / grouped5['Position_based_on_clustering'].count() * 100

print("The models accuracy: " + str(accuracy) + " %")
```

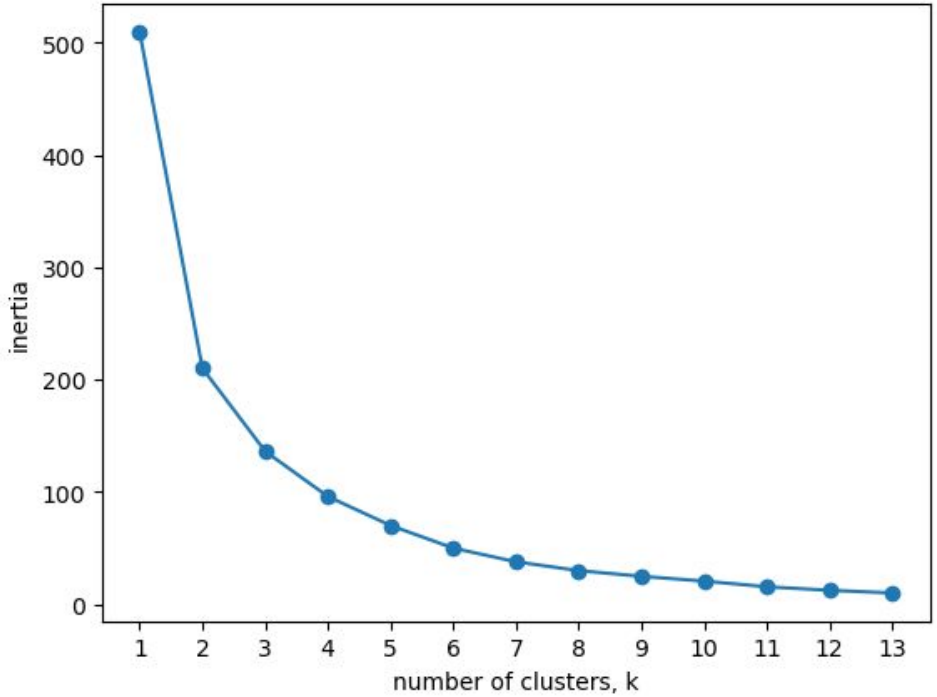
```
The models accuracy: 52.311435523114355 %
```

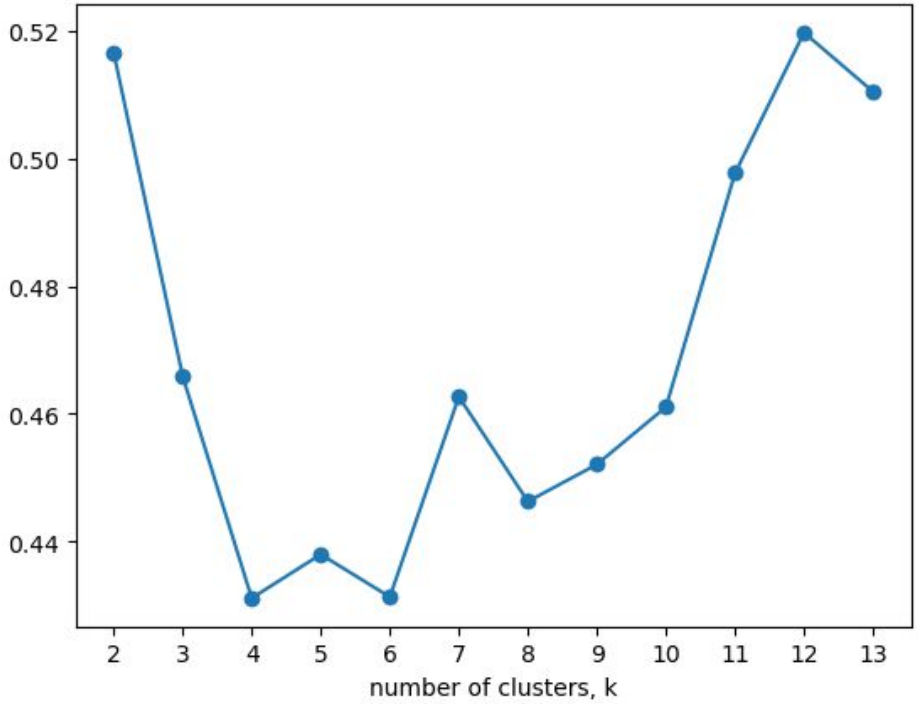
```
#Centers

center = grouped5[grouped5['POSITION'] == 'Center']
true = center[center['Check similarity'] == True]
print('Model accuracy: ' + str(true['PLAYER_ID'].count() / center['PLAYER_ID'].count() * 100) + ' %')
```

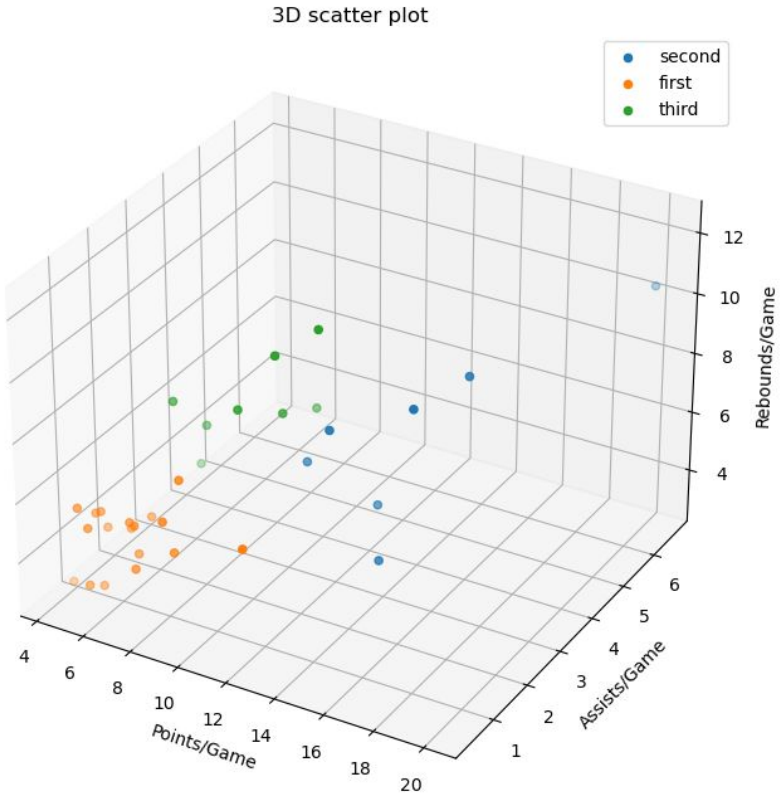
```
Model accuracy: 78.78787878787878 %
```

# PCA K-Means Clustering only centers



# 3D scatter plot of the clustered centers by P/G, A/G and R/G



# Conclusion

It was not possible to choose a clustering technique with which, based on the selected data, the players would have been relatively accurately assigned to their actual position by the machine learning model.
This also proves that in today's NBA, the different positions are often not so separated from each other and there are cases when, for example, a center has stats that are not traditionally typical for that position at all. The best example for this is Nikola Jokic:

| NAME | PTS_G | AST_G | REB_G |
|---|---|---|---|
| Nikola Jokic | 20.122511 | 6.657751 | 10.437264 |

For further developments, i could include more data points into the model, for example advanced stats.