Project Report

GitHub URL

https://github.com/eimhincullen/UCDAPA_eimhincullen

Abstract

The aim of this project is to segment PGA Tour golfers into clusters using their strokes gained data (Strokes gained: How it works, 2016) from the 2021/2022 PGA Tour season up to the 23/05/2022. Strokes gained measures how the player is performing versus the field across four different aspects of the game. The second aim of this project is to generate insights into these clusters using the PGA Tour's FedEx Cup rankings.

The steps to this project can be split into the following five sections:

1. **Importing Data:** This involves scraping the FedEx Cup rankings from the PGA Tour website and importing strokes gained data from a CSV file.
2. **Merging Data:** This involves reformatting a column and merging the two datasets together.
3. **Preliminary Analysis:** This involves reviewing data completeness, filtering the data and exploring the relationship between the variables in the dataset.
4. **Machine Learning:** This involves carrying out K-Means cluster analysis using the standardised strokes gained variables as the features. The clustering process is repeated with different combinations of features to optimise cluster quality. Hyperparameter tuning is used to identify the optimal number of clusters.
5. **Analysis and Insights:** The clusters are visualised to help understand the unique characteristics of each. The cluster each player belongs to is added to the dataset and further insights are identified.

Introduction

Golf is a game where people of all ages, shapes and sizes play side by side. Even at the elite level of the PGA Tour, players in their early twenties compete against veterans in their forties and fifties. The aim of the game is to get the ball in the hole is as few strokes as possible. Each player has different approaches to this depending on the strengths and weaknesses of their game.

Strokes gained is a great way to measure players performance versus their competitors across the four key aspects of golf: driving off the tee, irons on approach shots to the green, chipping around the green and putting the ball in the hole. Most PGA players monitor these stats closely to see where their strengths lie and what areas they need to improve.

Strokes gained is not intuitive, especially when looking across four separate categories for 150+ players. The aim of clustering strokes gained data is to identify groups of players with similar skill sets. By then analysing each cluster using the traditional metrics of success (wins, top 10's, FedEx cup rankings) we can identify which cluster/skill set has the most success on the PGA Tour.

This analysis may be useful to golf analysts, sports betters, avid golf fans and even players themselves in identifying what separates the average PGA Tour professional versus the world's best.

Dataset

This project contains two data sources that are merged together during the project to form the working dataset. The two data sources are:

Strokes Gained Data: This is a csv file of the performance table on the datagolf website (Performance Table, 2022). It contains strokes gained data per player for the 2021/2022 season as of the 23/05/2022. The key fields are: *ott_raw* (strokes gained off the tee), *app_raw* (strokes gained approaching the green), *arg_raw* (strokes gained around the green), *putt_raw* (stokes gained putting). Having positive strokes gained means the player is performing better than tour average and negative strokes gained means the player is performing worse than tour average. This data source provides the strokes gained data that is used for clustering players in this project.

FedEx Cup Data: This data contains the 2021/2022 FedEx Cup standings that was scraped from the PGA Tour website on the 23/05/2022 (FedExCup - Official Standings | PGA TOUR, 2022). The FedEx Cup is the season long points race on the PGA Tour that began in September 2021 and ends in August 2022. Players earn points for every cut they make with more points earned the higher up the leaderboard they finish. The top 125 players at the end of the season retain their tour card for the next season while the top 30 gain exemptions into majors and invitationals next season. The key fields in this data source are FedEx Ranking, Events Played, Number of Wins and Number of Top 10s. This data source provides performance data of the PGA Tour players that can offer insights into the unique characteristics of the clusters we identify.

Implementation Process

Importing Data:

- Import BeautifulSoup and requests.
- Use requests.get to return the webpage for the FedEx Cup standings.
- Use BeautfiulSoup to return the HTML code from the FedEx Cup standings webpage. Inspect the output (note that this line is commented out on the jupyter notebook file).
- Find the table of the FedEx cup standings within the HTML code and inspect its contents (note that the line for inspecting the contents is commented out on the jupyter notebook file).
- Create an empty list for the table rows. Use a for loop to iterate each row with the html tag “tr” and append each to the table rows list. Filter out blank rows and only include rows with 9 items.
- Create an empty list for each of the columns from the FedEx Cup standings we are interested in. Use a for loop to iterate over each table row and extract the relevant data for each player into each list.
- Import re and use a regex to remove html tags from player_name. Use replace to remove unnecessary text.
- Create a function called remove_tags that removes html tags and converts the data to integer. This function uses regex and the replace function. Apply this function to the numeric lists pulled from Fedex Cup standings two steps previously.
- Import pandas. Save the lists to a dictionary and convert to a csv called fedex_cup_ranking.csv. This step is necessary as the Fedex Cup ranking is a live table that updates after every PGA Tour event.
- Import a date stamped fedex_cup_ranking csv file from the 23/05/2022 and save as fedex_cup.
- Import a csv file containing the strokes gained data from datagolf as of the 23/05/2022 and save as strokes_gained.
- Print the head of the two data frames imported.

Merging Data:

- Extract the first name and surname of each player from player_name in the strokes_gained data frame and save to two new columns. Merge the two new columns and overwrite the previous player_name column. Drop the first_name and surname columns you just created. Print the head of the strokes_gained data frame to ensure player_name reformatted as intended.
- Merge the fedex_cup and strokes_gained data frames together using an inner join. Join on Player Name from fedex_cup and player_name from strokes gained. Merging on name is not ideal as there are some variations in the spelling of player names between the two data sources. For example, Matthew Fitzpatrick is down as Matt Fitzpatrick in the fedex_cup data frame. These players have been dropped from the data frame. This is a small number of cases and there is still enough players for analysis and clustering. Save the merged data frame as df.
- Drop unnecessary columns from df and print the data frame to inspect it.

Preliminary Analysis:

- Check for any null values in the dataset. Since there are none, there is no need to drop null values or replace nulls. This was helped by the data cleaning when web scraping and using an inner join when merging.
- Use .describe to summarise the numerical data in the data frame.
- Import matplotlib, seaborn and numpy. Create a histogram to get the distribution of the number of events played by players in the data frame. Use a numpy array to set the bins of the histogram.
- Remove any players who have played less than 10 events from the data frame. This is to ensure players have enough events played to provide accurate strokes gained data. Use .describe to summarise the numerical data in the updated data frame. This resulted in a 12% reduction in the data frame but still 199 rows of data
- Save the 6 strokes gained columns in a list called df_sg. Create a pairplot to explore the relationship between the 6 strokes gained variables.

Machine Learning:

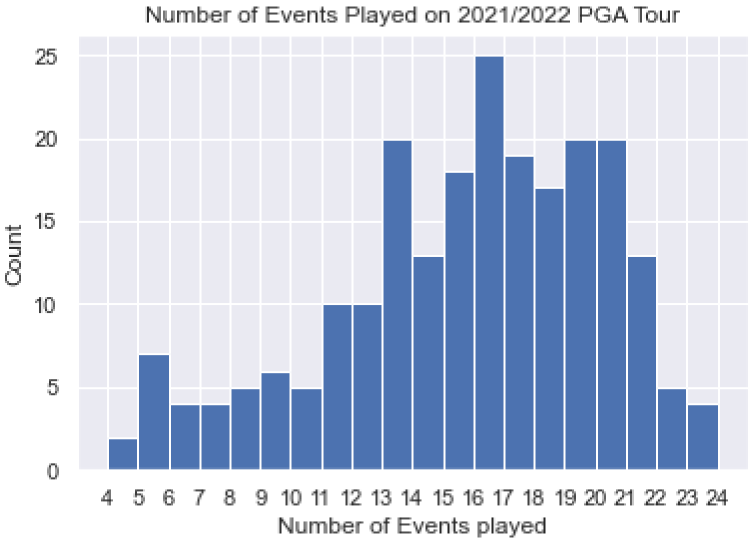
- Import KMeans, MinMaxScaler, KElbowVisualizer and warnings. Ignore warnings.
- Save the four strokes gained variables (putt_raw, arg_raw, app_raw, ott_raw) in a list features.
- Instantiate MinMaxScaler and scale the features list. Save the scaled features as a data frame. Use MinMaxScaler as strokes gained values can be both positive and negative.
- Instantiate KMeans and the KElbowVisualizer across a numpy array 1 to 11. Fit the scaled features to the visualizer. Show the visualizer to get the elbow plot and identify the optimal number of clusters. This is an example of hyperparameter tuning. As we are using an unsupervised learning technique boosting is not applicable.
- Create the KMeans model with the optimal number of clusters and fit the model to the scaled features. Calculate the inertia of the model to measure quality of clusters.
- Repeat the last four steps four times dropping one of the strokes gained variables each time. Find the optimal number of clusters, fit the model and calculate the inertia. In the end the model with the scaled features putt_raw, app_raw, ott_raw has the lowest inertia value and is the chosen clustering model.

Analysis and Insights:

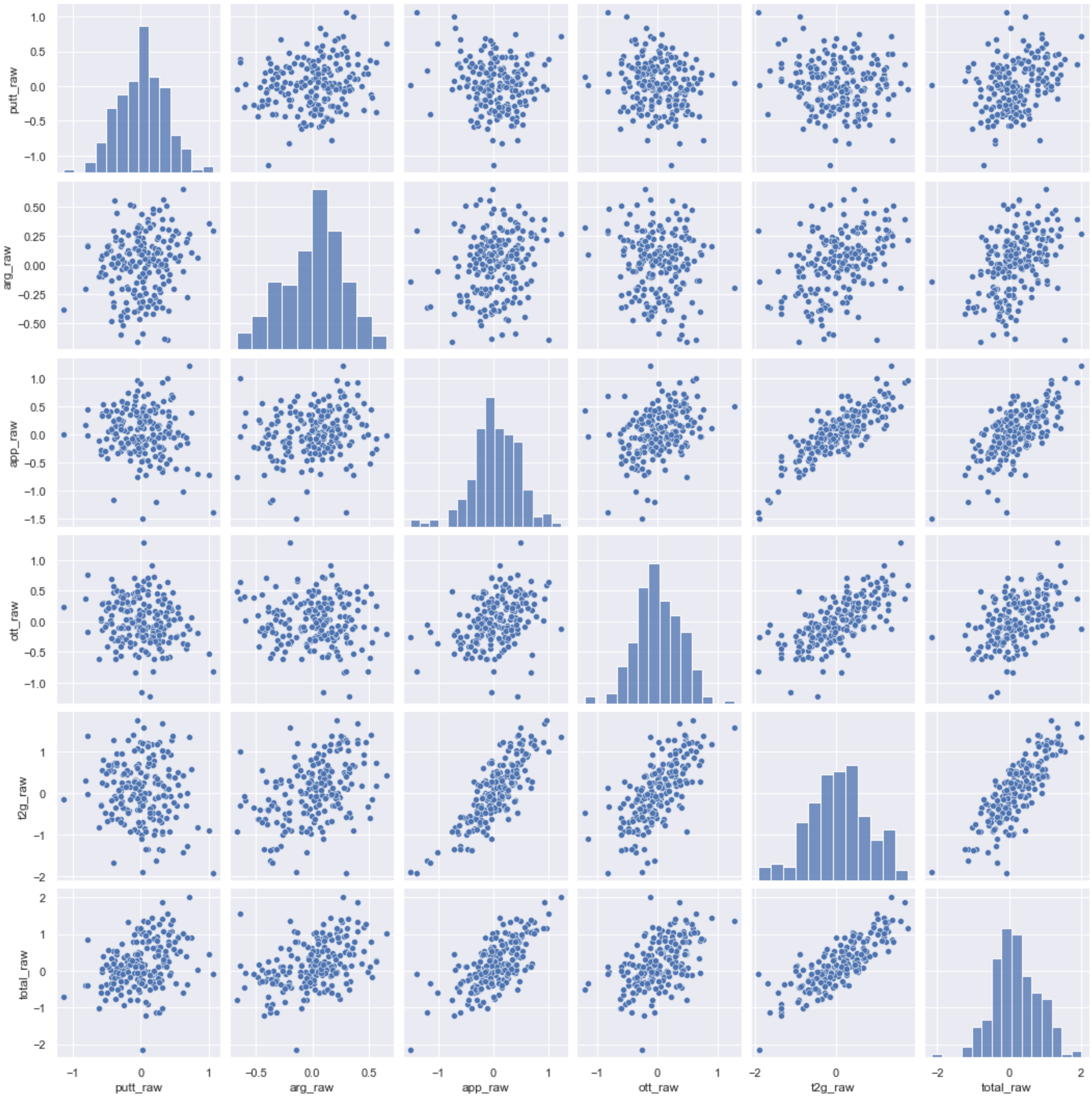
- Import Axes3D from mpl_toolkits. Create a 3d plot of the 4 clusters identified using the scaled features putt_raw, app_raw, ott_raw. The scaled strokes gained putting are on the x-axis, the scaled strokes gained approach are on the y-axis and the scaled strokes gained off the tee are the z-axis.
- Add the cluster labels to a new column in the df data frame called cluster and print the head of your data frame.
- Get the value counts of each cluster in your data frame.
- Calculate the mean and median per cluster across the numerical variables in the data frame.
- Add two new columns to the data frame, one that flags players inside the top 30 of the FedEx Cup rankings and one that flags players outside the top 125 of the FedEx Cup rankings. Calculate the number and percentage inside the top 30 per cluster and calculate the number and percentage outside the top 125 per cluster.
- Add a boxplot of FedEx Ranking per Cluster.
- Find the number of wins and number of top 10s per cluster.
- Filter the dataset to find the one player outside the top 125 in Cluster 0.

Results

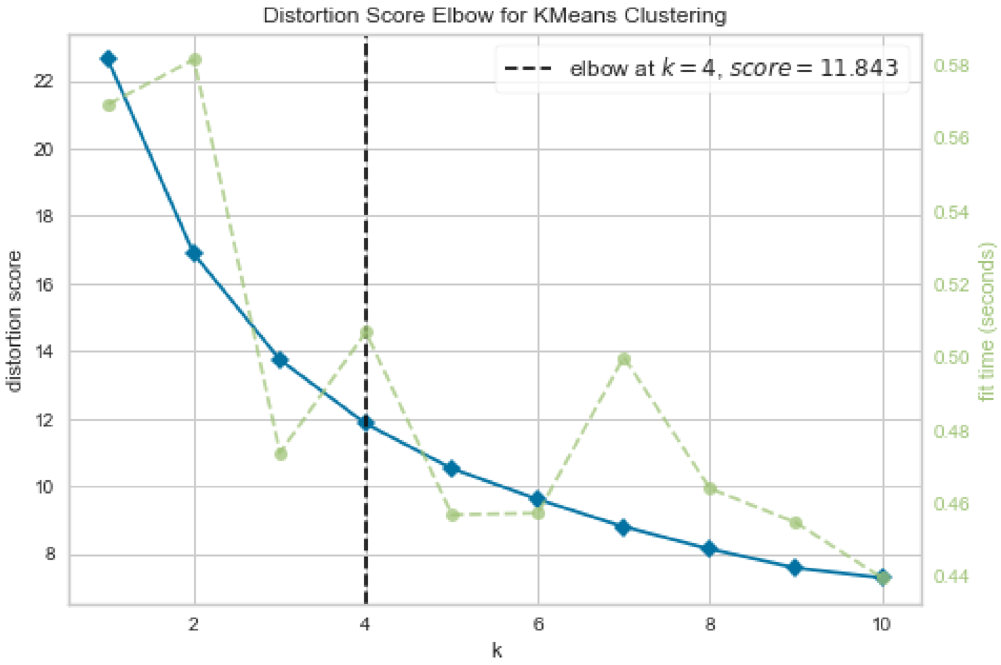
The first graph below shows a histogram of the number of events PGA Tour players had played in the 2021/2022 season up to 23/05/2022. Values ranged from 4 to 24 events. Any players who had played less than 10 events were removed from the dataset to avoid having too few samples to accurately calculate their strokes gained data. This reduced the dataset by 12% from 227 players to 199.



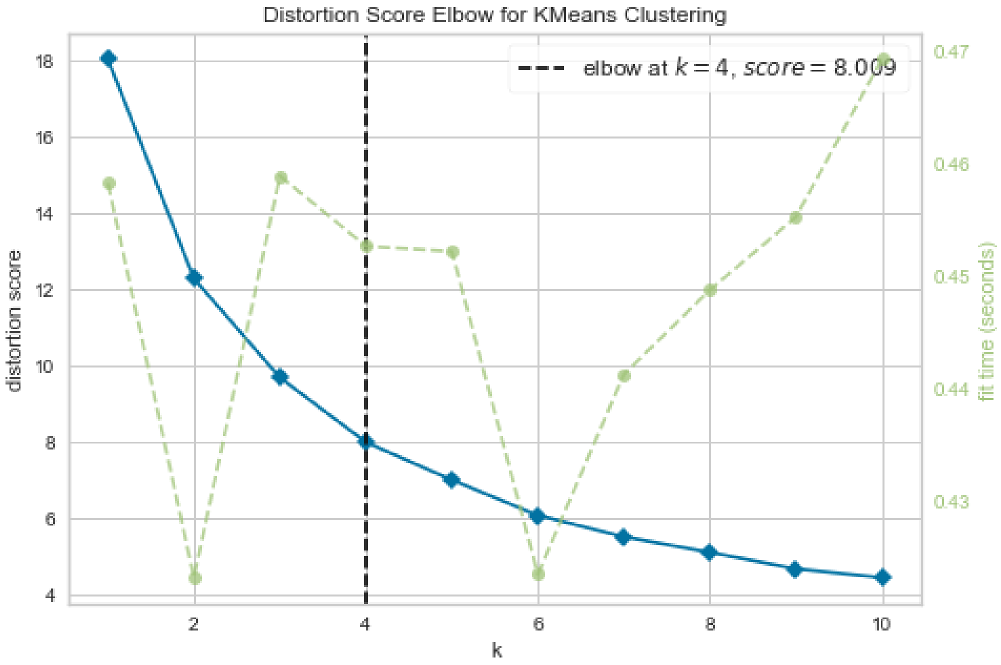
The graph below shows a pairplot for the 6 strokes gained variables in the dataset. This shows a scatterplot for each variable against the other 5 variables. Note that there is no correlation between putt_raw, arg_raw, app_raw and ott_raw. T2g_raw and total_raw had a positive correlation with each other and somewhat of a positive correlation with app_raw and ott_raw. This makes sense as t2g_raw and total_raw are combinations of the other strokes gained variables. These two variables were not included in the features when clustering as a result.



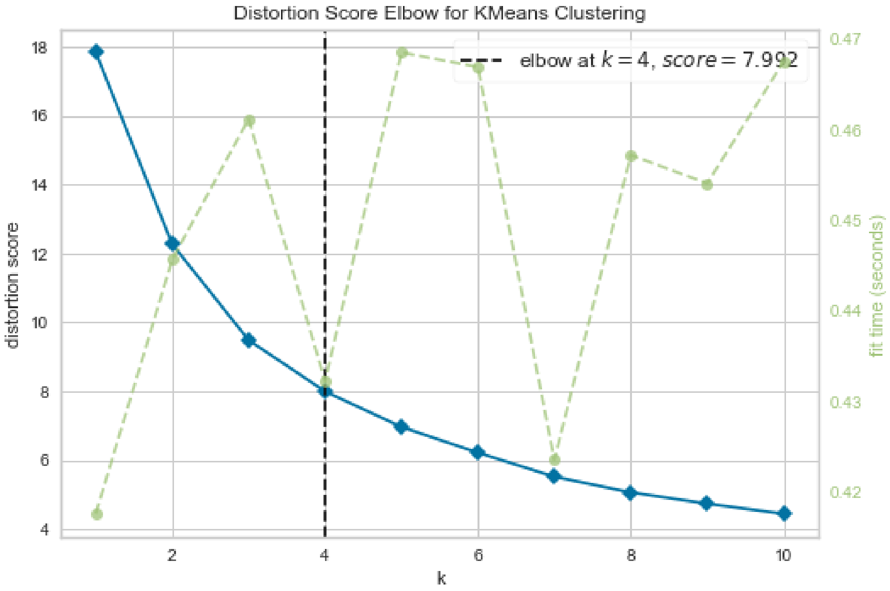
The graph below shows the Elbow Plot for K-Means Clustering with the 4 strokes gained (putt_raw, arg_raw, app_raw and ott_raw) variables. The optimal number of clusters is 4 and the inertia score is 11.843. K-Means Clustering was repeated with the combinations of 3 of the 4 variables to see which achieved the lowest inertia score.



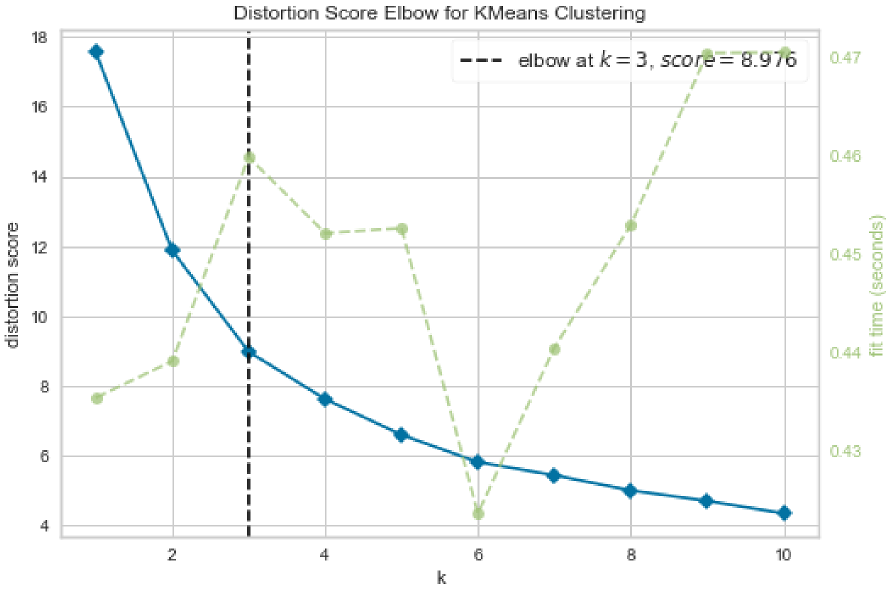
The graph below shows the Elbow Plot for K-Means Clustering with the 3 strokes gained: putt_raw, arg_raw and app_raw variables. The optimal number of clusters is 4 and the inertia score is 8.009.



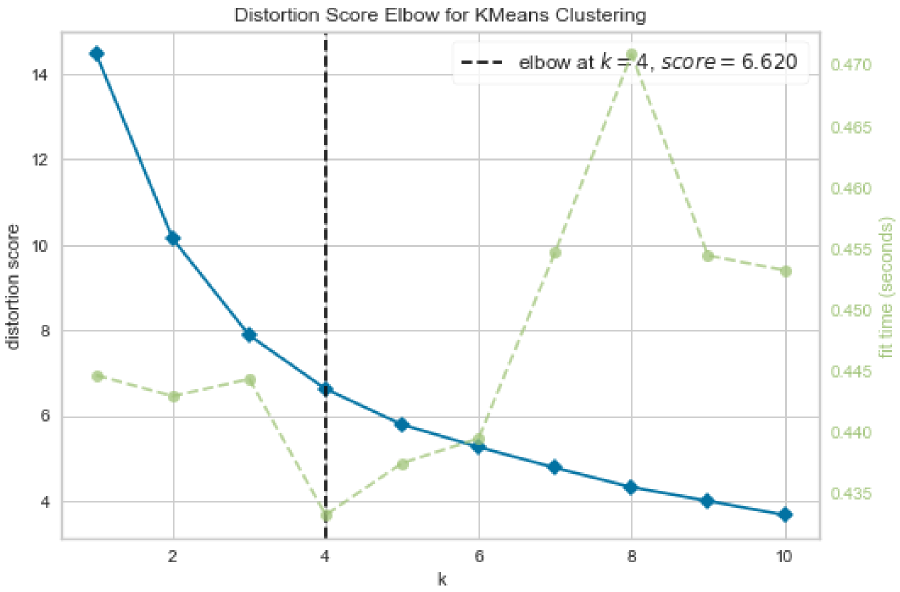
The graph below shows the Elbow Plot for K-Means Clustering with the 3 strokes gained: putt_raw, arg_raw and ott_raw variables. The optimal number of clusters is 4 and the inertia score is 7.992.



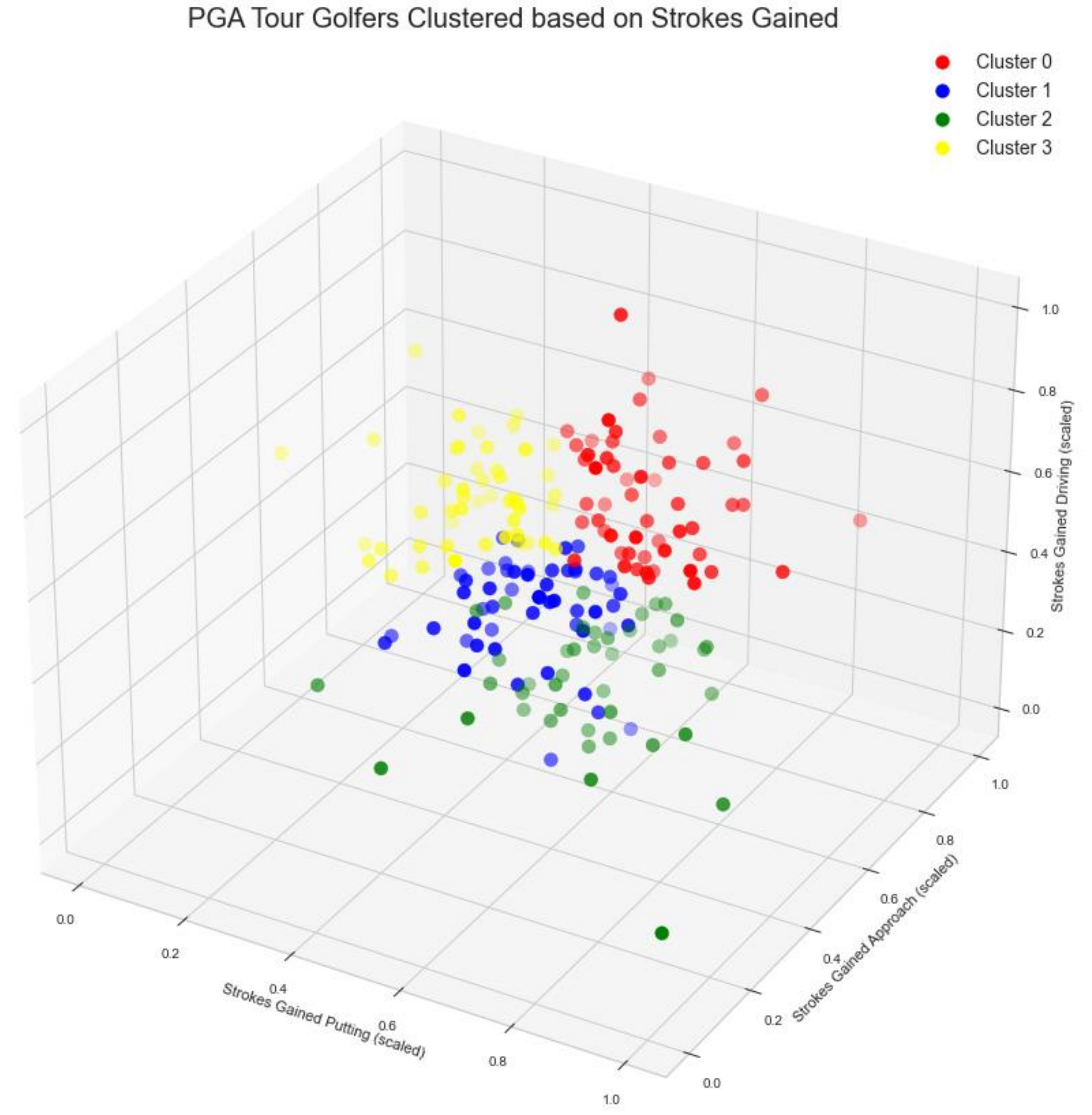
The graph below shows the Elbow Plot for K-Means Clustering with the 3 strokes gained: app_raw, arg_raw and ott_raw variables. The optimal number of clusters is 3 and the inertia score is 8.976.



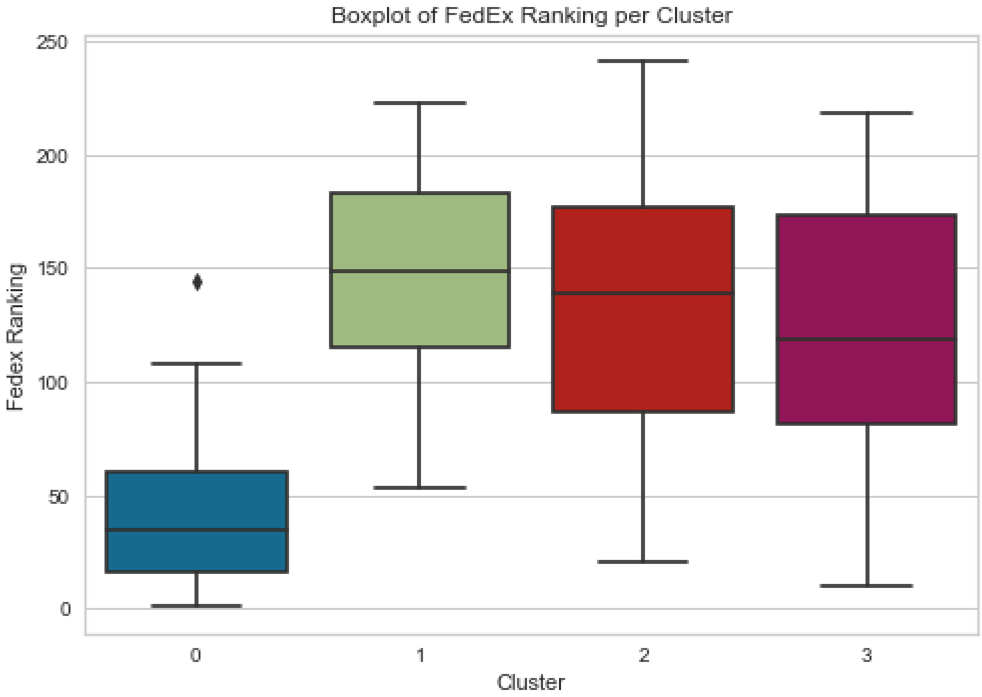
The graph below shows the Elbow Plot for K-Means Clustering with the 3 strokes gained: putt_raw, app_raw and ott_raw variables. The optimal number of clusters is 4 and the inertia score is 6.620. This is the model with the lowest inertia score and the model chosen to visualise and identify insights.



The graph below visually represents the clusters using putt_raw, app_raw and ott_raw as features. Cluster 0 and Cluster 3 are quite distinctive while there is overlap between Cluster 1 and Cluster 2. Players in Cluster 0 appear to perform well in all three strokes gained categories particularly approach and somewhat driving while players in Cluster 1 appear to struggle off the tee and are in the bottom half of putting. Cluster 2 seems similar to Cluster 1 except players tend to perform better in strokes gained putting. Cluster 3 who perform well in strokes gained approach but perform poorly in strokes gained putting.



The graph below shows a boxplot of the players FedEx Cup ranking per cluster. Cluster 0 is the best performing cluster with the lowest median ranking and the smallest spread. There is not much difference in the median value for the 3 remaining clusters. Cluster 3 has the next lowest median. Cluster 1 has the highest median and a smaller spread than Cluster 2 and Cluster 3.



Insights

- Cluster 0 is the best performing cluster out of the 4 identified across all variables that measure success on the course. They account for 82% of the wins on tour this season. This group of players have on average 3.32 top 10s while the other 3 clusters average just above or below 1 top 10.
- Cluster 1 is the worst performing cluster on average. They have the lowest average FedEx Cup ranking, lowest number of wins and top 10's. They are averaging -0.314 strokes gained off the tee as a group and this seems to be holding them back the most.
- The similarity seen in the cluster visualisation between Cluster 2 and Cluster 3 translates to the performance metrics also. They are averaging nearly an identical number of wins and top 10s. Cluster 2 struggles off the tee and with approach shots while Cluster 3 struggles with putting.
- Players in the top 30 of the FedEx Cup rankings at the end of the season receive huge cash bonuses and exemptions into elite tournaments next season. 24 out of the 27 players from the top 30 in the dataset belong to Cluster 0. No players from Cluster 1 are in the top 30.
- Players outside the top 125 of the FedEx Cup rankings at the end of the season lose their full playing privileges next season. Some will drop down to the tour below. 68% of the players in Cluster 1 are currently outside the top 125. One player from Cluster 0 is outside the top 125. This player is Austin Smotherman. He is above average for Cluster 0 in strokes gained off the tee and approach, however he is well below average for strokes gained putting and around the green.

References

PGATour. 2016. Strokes gained: How it works. [online] Available at: <https://www.pgatour.com/news/2016/05/31/strokes-gained-defined.html> [Accessed 28 May 2022].

Datagolf.com. 2022. Performance Table. [online] Available at: <https://datagolf.com/performance-table> [Accessed 28 May 2022].

PGATour. 2022. FedExCup - Official Standings | PGA TOUR. [online] Available at: <https://www.pgatour.com/fedexcup/official-standings.html>