COMP3021 Fundamentals Information Visualisation

Analysis of PGA Tour Golf Data (2015 – 2022)

Contents

Dataset

Dataset pulled from - https://www.advancedsportsanalytics.com/pga-raw-data

Description Of Data & Background Information

This dataset is a collection of information collected from all tournaments that have taken place on the PGA tour across the 2015 – 2022 seasons. The PGA tour is considered the pinnacle of the professional golf tours making it the most suitable to analyse the performance of the elite level players.

This data contains the following features:

- Player_inital_last -> player name in the format (first name initial).(surname)
- tournamentID – identifier for the tournament
- hole_par -> total par of the holes played
- strokes -> strokes made by player
- hole_dkp -> unused variable
- hole_FDP -> unused variable
- hole_sdp -> unused variable
- streak_dkp -> unused variable
- streak_FDP -> unused variable
- streak_sdp -> unused variable
- number of rounds -> number of rounds played per tournament
- made_cut -> whether player made cut
- total_dkp -> unused variable
- total_fdp -> unused variable
- total_sdp -> unused variable
- player -> Player full name
- tournament -> Tournament name
- course -> Course name
- purse -> Pot of prize money (million)
- season -> Season played
- finish -> Finishing position
- sg_putt -> Strokes gained when putting
- sg_arg -> Strokes gained around the greens (chipping/putting)
- sg_ott -> Stokes gained off the tee (tee shots)
- sg_t2g -> Strokes gained tee to green
- sg_total -> Total strokes gained

Understanding Golf/Data/Problem

To understand the features and analysis a basic understanding of the game of golf is required.

Objective:

The aim of golf, in the most common tournament format, stroke play, is to achieve the lowest score by taking the fewest amount of stokes. In the PGA tour all participants of a tournament play on Thursday and Friday, 36 holes total; the top 70 (all players matching the 70th players score or better) advance to play an additional 36 holes on the Saturday and Sunday to compete for the tournament. This is called the cut and if a player does not make the cut, they do not receive any winnings. There are 47 official PGA tour tournaments per season. Players on the tour are not required to play every tournament however must play at least 15 per season. Different tournaments have different size fields (50-150 players).

Scoring:

Each hole has a par, this is the number of strokes a player is expected to complete said hole. After each hole the player is given a score based on the number of strokes made on that hole in relation to its par.

X → Number of strokes (as a player can go as far over par as possible)

[(score name, effect on player score, relation to par of hole)]

The system goes as: [(Condor, -4, 4 shots under par), (Albatross, -3, 3 shots under par), (Eagle, -2, 2 shots under par), (Birdie, -1, 1 shot under par), (par, 0, level with par), (Bogey, +1, 1 shot over par), (Double Bogey, +2, 2 shots over par), (Triple Bogey, +3, 3 shots over par), (*x* Bogey, +*x*, *x* shots over par)]

For example: player who makes 4 stokes on a par 3 has achieved a bogie and +1 is added to their overall score.

The total score of a player is then calculated by adding all these strokes taken per hole together, for all the holes completed on the round (typically 18/day or round, 72/tournament). This will compare with the par of the course and the score is the difference between the strokes and par. The lower the score the better, therefore players aim to get their score as negative as possible in contrast to other sports.

Strokes Gained (SG):

This report will often reference the strokes gained metric. Strokes gained is a statistical measure used to evaluate a golfer's performance compared to the expected performance of a golfer of the same level. This can be their overall performance or for a specific part of their game (putting, chipping, etc…). It is calculated by subtracting the expected number of strokes (taken from historical data of PGA tour golfers, which is then adjusted based on shot/hole specific factors) from the actual strokes taken. A positive SG indicates performing above the average (gaining strokes) and a negative SG indicates performing below the average (losing strokes).

Figure 1: Infographic Explaining SG 9ft Putt Example

Data Cleaning

Dataset transformed to remove all but the data interest:

removed attributes [Player_initial_last, tournament id, player id , hole_DKP, hole_FDP, hole_SDP, streak_DKP, streak_FDP, streak_SDP, finish_DKP, finish_FDP, finish_SDP, total_DKP, total_FDP, total_SDP, Unnamed: 2, Unnamed: 3, Unnamed: 4, tournament name]

Initial Questions

Each question and visualisation is described and critically discussed.

<Q1> Is Strokes Gained a Good Metric for Evaluating Performance?

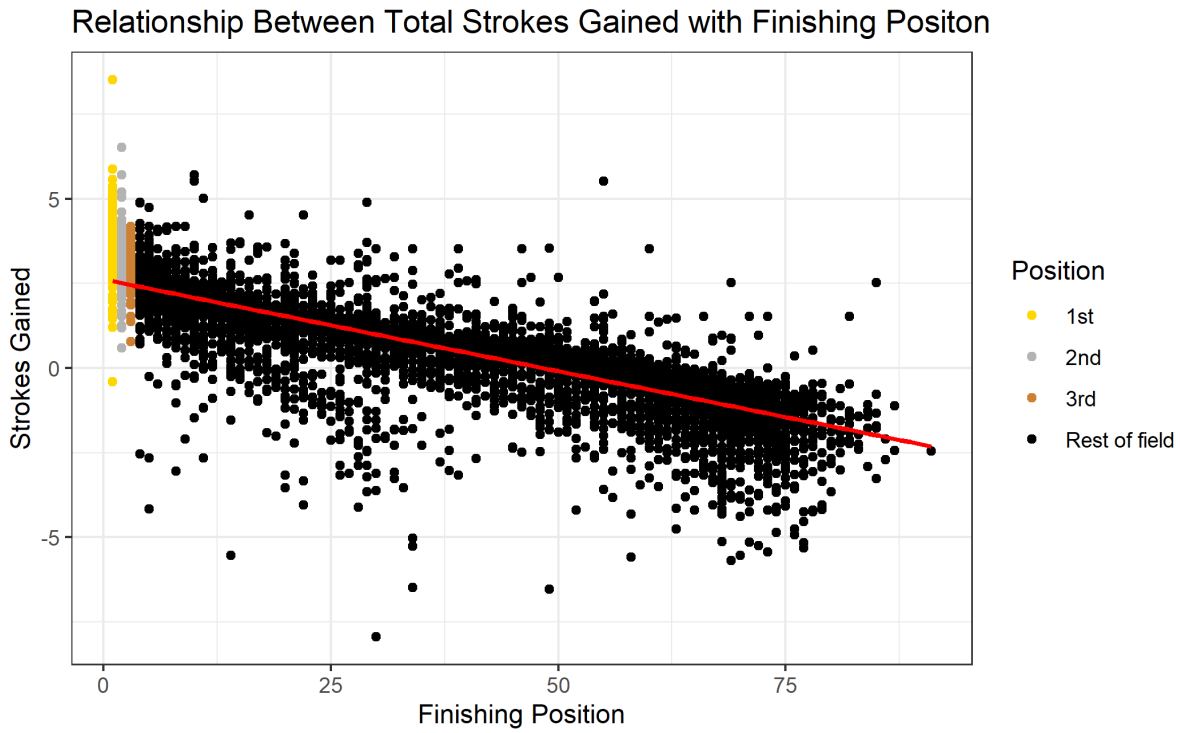

Figure 2: Relationship Between Total Strokes Gained with Finishing Performance

As discussed, strokes gained is a commonly used metric used to evaluate a golfer's performance. The goal of this question and visualisation is to merit this common assumption. Given the objective of this visualisation is to compare the relationship between two variables (continuous and non-continuous) a scatter plot has been used.

Cleaning the data for this plot involved removing missing values and outliers using 'na.omit' and 'filter'.

To make the 1st, 2nd, and 3rd visually distinguishable to the rest of the of the finishing positions their colour has been encoded to gold, silver, and bronze. This was done throught the use of case_when(). The use of distinct and recognisable colours aids in the expressiveness of this plot as it highlights the importance of top finishing positions quickly to the viewer. The colour choices have been chosen as they are commonly associated with these positions.

A trend line has also been encoded, equation y ~ x, to aid the viewer see the overall pattern of the relationship between strokes gained and finishing position. Expressing the trend in the data much more clearly. It has been coloured red to visually stand out to the viewer.

Effective labels have been used for the axis and legend.

From this plot we can see that generally if better finishing positions tend to gain more strokes. Suggesting that strokes gained is a good metric for evaluating performance.

<Q2> Does a larger purse size result in better player performance?

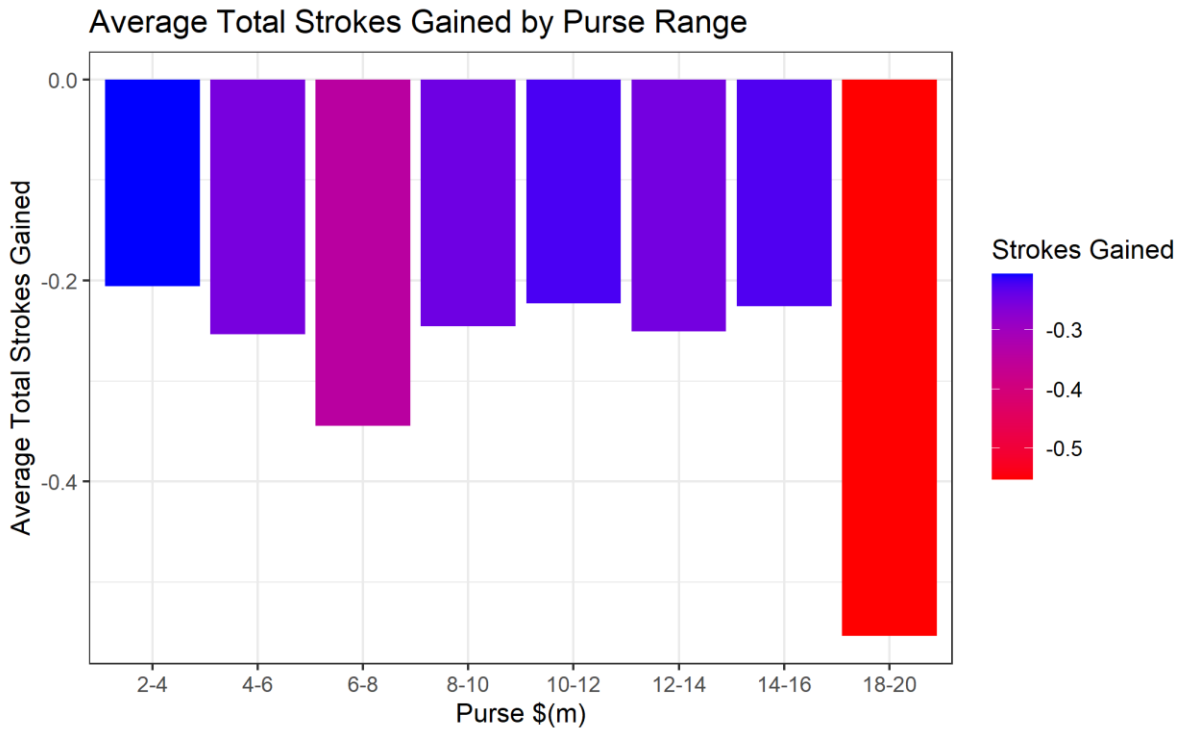

Figure 3: Average Total Strokes Gained by Purse Range

For the second question, visualisation has been used to investigate if a larger purse size resulted in better performance. This questions was derived from the thought process that if there was more monetary incentive to win, would players perform better?

The purse sizes were categorized to aid simplify the data, to make it easier to interpret. Then for each of these categories the average total strokes gained was calculated.

This meant it was possible to use a bar plot with the x-axis representing the purse size and the y-axis the average total strokes gained. To emphasize the comparison the fill was encoded with the total strokes gained in a gradient scale, blue to red. Where blue represented the better performance (higher strokes gained) and red worse performance. This follows Mackinlay's recommendation, from (*Mackinlay, 1986*). to use colour gradients to represent quantitative variables.

From this plot we can see that for the majority of the purse ranges, 2-18, there is no major change in average strokes gained, this is effectively visualised by the similar size of the bars and then emphasised by their similarity in colour. However, there is a drastic drop off in performance at the 18-20 purse size range, emphasised by the strong contrast in colour and size to the rest of the bars. Suggesting that in the highest stake tournaments, performance actually decreases.

<Q3> Which seasons did players score the best?

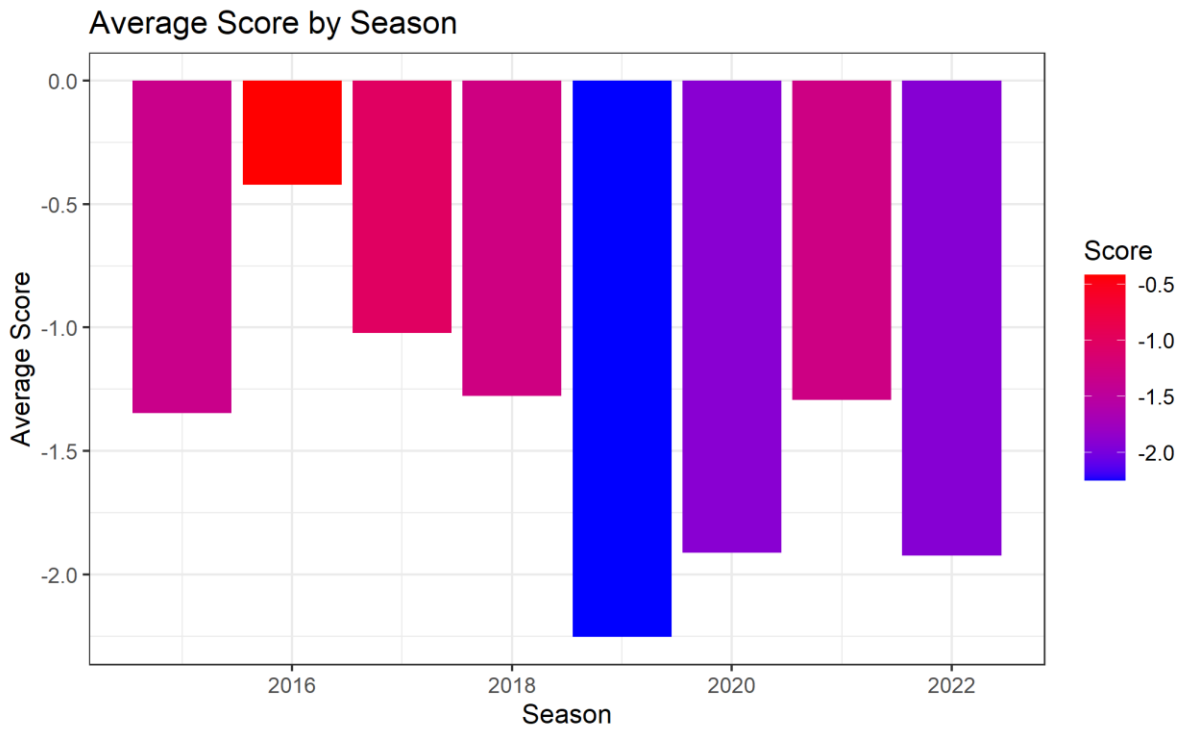

Figure 4: Mean Score by Season

To discover which season players scored the best, the visualisation had to illustrate a comparison by the average score between seasons.

The data set did not already contain a score variable therefore a new “score” variable was created by subtracting the number of strokes taken from the par of each hole (reference the background information, scoring for reasoning). The scores were then grouped into seasons and the seasons average score was calculated, using the group_by() to group seasons and summarize() to calculate the average.

This data was encoded in a bar plot where each bar represents the average score per season. Gradient fill was used to improve emphasis by encoding the size of the average score, blue representing better scoring seasons and red worse scoring seasons.

From this plot it is clear there is no real trend to investigate further, but there is a clear worst scoring season on average, 2016 and the 2017 season was the best scoring on average.

<Q4> What is the proportion of players who scored under par, even par and over par?

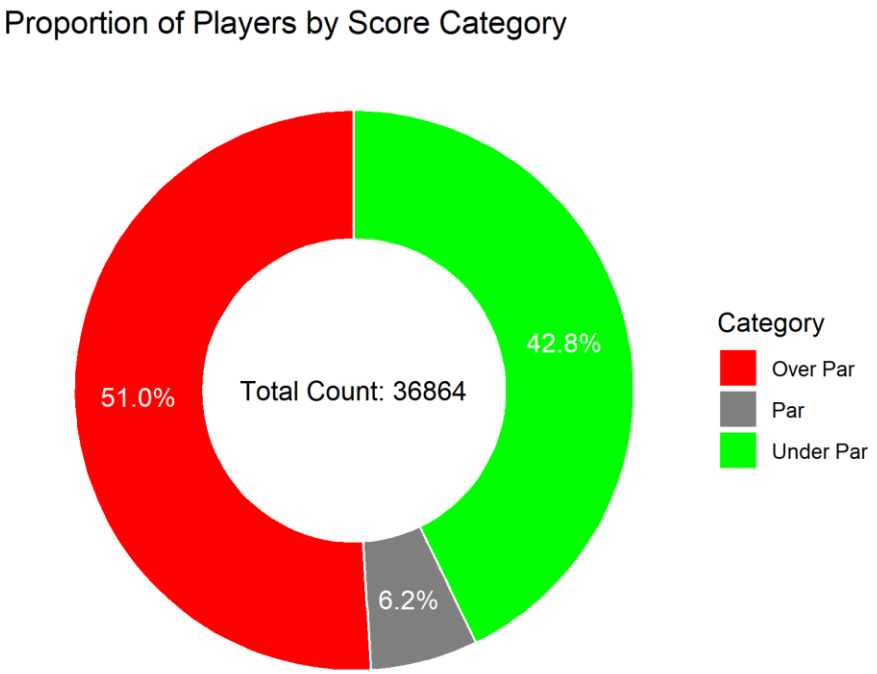

Figure 5: Proportion of Players by Score Category

Question 4 aims to determine the proportion of players who scored under par, even par and over par of all the players scores across the seasons 2015-2022.

The score variable created for the visualisation above was used and compared to 0 to determine its category. Case_when() was used to create the new variable storing the category of the score. The categories were then summed to store the number of scores for the respective categories. An additional variable was then created in this new dataset to store the percentage. Using the mutate function percentage was created and calculated as follows: percentage = Count / sum(Count) * 100. These calculations included the use of other dplyr functions such as count(), rename() and mutate()

To then visualize the proportion of players in each category a pie/donut chart was used as this questions objective is simply to visualize the composition of a static variables share of a total. In terms of encoding, the colours chosen were chosen for their easy interpretation, where green represented Under Par, grey represented Par, and red represented Over Par. The addition of the percentage labels made it easier for the viewer to grasp the exact proportion of players in each category.

Further Questions

<FQ1> Which is the most important part of the game to gain strokes to win tournaments?

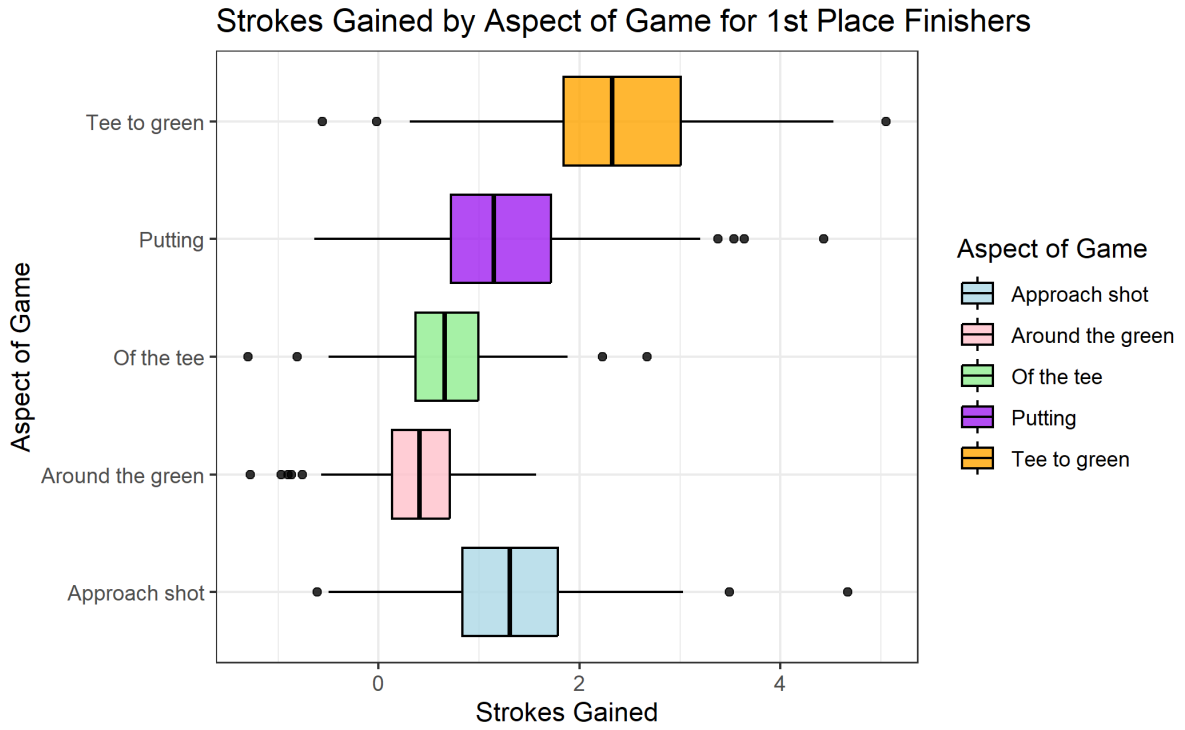

Figure 6: Strokes Gained by Aspect of Game for 1st Place Finishers

Having established from <q1> that strokes gained in a good measure of performance, there is now confidence in using it in more depth when evaluating players. This allowed us to further explore the affects of how each aspect of the game is the most important when competing for tournaments. Delivering the question, which is the most important part of the game to gain strokes to win tournaments.

To answer this question, the data needs to be filtered to only include players who finished in 1st place. The subset() function is used to create a new data frame with only these players. From these players only the strokes gained for each aspect were collected into a new data frame, where gather was then used to reshape the df so the value and sg_aspect were stored in two singular columns.

A boxplot was the suitable encoding as this questions requires the viewer to have to compare data from different categories. This visualisation graphically depicts the different strokes gained groups through their quartiles, whilst still highlighting mean, and outliers. Coord_flip() was used to improve the effectiveness of the plot as by flipping the plot horizontally the boxes and the difference between them are more readily perceived/easier to visualize. The boxes are encoded to different colours per aspect helping viewers distinguish each box.

The plot shows that the most important aspect of the golf game to gain strokes is from “tee to green” for tournament champions. As despite a few outliers visualised by the left whiskers, all values (lines in the box) are further along than the other aspects.

<FQ2> How do strokes gained in different shot categories compare between tournaments with the highest purse range and tournaments with other purse ranges?

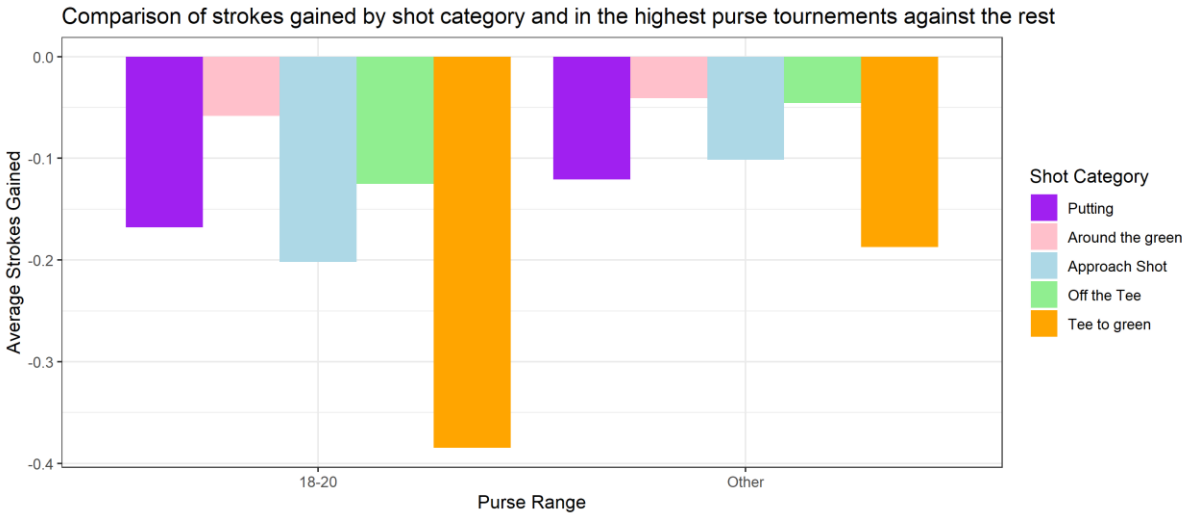

Figure 7: Comparison of Strokes Gained by Shot Category and in the Highest Purse Tournaments Against the Rest

From the results of <Q2> it was clear that there was a significant drop of in performance in the highest purse tournaments. Given other findings in the study supporting strokes gained as a metric of performance, its now possible to confidently explore what aspect of a player's game is affected the most in these high-pressure tournaments causing this drop off in performance.

To transform the data, first any rows where strokes gained data was missing for any of the shot categories were filtered out. "purse_range_new" was a new variable created to group the tournaments into two categories - those with the highest purse range ("18-20") and those with other purse ranges. We then calculated the mean strokes gained for each shot category in each purse range group using the group_by() and summarize() functions.

A dodged bar plot was used to visualise the data as this plot allows the viewer to compare the average strokes gained for each shot category within the two different purse range groups side by side. Making it easy for the viewer to compare performance. In terms of encoding, the y-axis displays the average strokes gained and the x-axis the shot category. Each bar has been coloured dependent on the shot category. These colours are distinct, improving effectiveness and are the same colours as the previous plot that visualised the same categories, making use of associated colours.

From this plot it is clear that in the high purse tournaments there is not a huge decrease in all the shot categories except for the tee to green. Suggesting in these large purse tournaments something is affecting this aspect of the game negatively.

Reflection

This project aimed to explore a dataset and through the use of R transform and encode the information so that a set of questions could be answered.

Initially it was difficult to settle on a suitable dataset as the datasets available from commonly used repositories such as: Rdatasets & UCL ML repo, were either not suitable (dimensionality issues) or did not create enough interest to be analysed. However given my external interests and the ever-expanding world of sports statistical analysis, I was able to source the dataset used in this project,

PGA tour data from 2015-2022. Which not only created interest, but for the most part was suitable and also spanned for a long enough period of time to be considered a longitudinal dataset, meaning it can be confidently used to identify tends and patterns.

The initial questions were established through initial personal interest, as well as undergoing some initial data exploration once cleaned. From the plots made to answer these initial questions it was possible to then 're-frame' any initial conceptions about the data and ask to follow up/further questions to explore the results of these initial questions.

Some key findings in this project were that strokes gained is a suitable metric for performance and the most important aspect of a game for a tour golfer is tee to green. One limitation of this project and its dataset is the lack of data on the specific golf courses. Its clear that in some seasons players performed better/worse, having additional information on the types of course played per season would have enabled further exploration on why these seasons average scores differed and if it related to the courses played that year.

Overall, the plots/visualisations of this project clearly have effectiveness and expressiveness at the forefront of its purpose. Suitably picking encoding methods for each questions allows viewers to be able to see true data in a way that is more readily perceived.

References

Figure 1, Infographic Explaining SG 9ft Putt Example, (https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.nationalclubgolfer.com%2Fnews%2Fstrokes-gained-explained-2%2F&psig=AOvVaw2RrUcF4IYz11xOzw-cXklD&ust=1681848025315000&source=images&cd=vfe&ved=0CBAQjRxqFwoTCJjkoqPasf4CFQAAAAAdAAAAABB6)

Mackinlay, Automating the design of graphical presentations of relational information, 1986.