**DC Project**

# GitHub URL

URL: <https://github.com/castolo9/UCDPA_Darragh-Claffey>

# Abstract

Statistics are very important aspect of golf in gauging performance and in recent years’ there has been a move away from traditional statistics to Strokes Gained (“SG”) statistics, developed by Mark Broadie.

For this assessment, the main goals were to:

* Analyse 2021 PGA Tour SG data to see how SG statistics in 2021 compared to those between 2004 and 2012; and
* Look at the two Ryder Cup teams (USA and Europe) in 2021 and see how they compared before the event.

# Introduction

Traditional golf statistics focused only on a players own performance, with measures things like:

* number of drives per round that landed in the fairway (“FIR”);
* number of greens per round hit in two shots under the par for each hole (“GIR”);
* total number of putts per round, etc.

The issue with these statistics is that they do not provide any insight into how the player performed in relation to other players in that event (e.g. on the PGA Tour) or other players of a similar standard (e.g. how a scratch golfer compared during their round with how other scratch golfers would have performed). If Tiger Woods hit 10 GIR and took 32 putts, did he perform well or not? SG statistics aim to provide a more detailed analysis of a players’ performance versus others.

As an example, tour pros on the PGA Tour took an average of 2 putts when 33 feet away from the pin (between 2004 and 2012). If Tiger Woods hits an approach shot and leaves the ball 33 feet from the pin and goes on to hole his first putt, he has gained one shot against the tour average. Using the traditional approach to statistics of counting the number of putts in a round, the 33 foot putt that Tiger holed counts as 1 shot and counts the same as if he had hit his approach to 1 foot and holed that putt. Therefore, SG seeks to quantify the difference in performance or difficulty of each shot – in this case, holing a 33 foot putt is more difficult than holing a 1 foot putt (the average is to take two putts from33 feet). (Broadie, 2014)

SG statistics use the following formula to analyse a player’s performance:

* **Strokes Gained Total (“SGT”) = Strokes Gained Tee-to-Green (“SGT2G”) + Strokes Gained Putting (“SGP”)**
  + **SGT:**
    - *“Compares a player's score to the field average. For example, a player will gain three strokes on the field if he shoots 69 on a day when the field averages 72. A player who shoots 74 on that day loses two strokes to the field”.* (PGA Tour, 2016)
  + **SGP:**
    - *“Measures how many strokes a player gains (or loses) on the greens”.* (PGA Tour, 2016)
    - *Between* 2004 and 2012, SGP contributed 35% to the SGT of all winners on the PGA Tour. (Broadie, 2014)
  + ***SGT2G:***
    - *“Measures all strokes not taken on the putting green”.* (PGA Tour, 2016)
    - *Between* 2004 and 2012, SGT2G contributed 65% to the SGT of all winners on the PGA Tour. (Broadie, 2014)
    - SGT2G can be further broken down into 3 areas:
      * SG Off-the-Tee;
      * SG Approach-the-Green; and
      * SG Around-the-Green.

(PGA Tour, 2016).

For the purposes of this assessment, SGT, SGP and SGT2G will be the primary focus.

# Dataset

The PGA Tour provide the facility to look up SG statistics on their website but do not provide the facility to download any of this data. As a result, would have been very time consuming to manually scrape this data and paste into an excel sheet as there were 44 different events during the 2021 season.

Therefore, used the following sources for data used in this assessment:

* 1. Advanced Sports Analytics (“ADA”) published a CSV file (“2021 PGA Tour Season Data”) on their website which included all SG statistics for the 2021 PGA Tour (data provided primarily for people playing online fantasy sports games like DraftKings); and
  2. Official World Golf Rankings (“OWGR”) as at the end of the 2021 PGA Tour season, manually scraped from PGA Tour website and pasted into an OWGR CSV file.

**Note:** Only manual changes made to the ADA CSV were to change some player names so that they matched OWGR (e.g. changing “Alexander Noren” to “Alex Noren”, etc.).

# Implementation Process

**Step 1:** Import all required packages and CSV’s:

1. Imported:
   * Pandas, Numpy, Matplotlib, Seaborn and OS;
   * 2021 PGA Tour Season Data CSV as a dataframe (“**pga2021**”); and
   * OWGR CSV as a dataframe (“**owgr**”).
2. Created new dataframe (“**owgr2021**”) which dropped certain unnecessary columns.
3. Created new dataframe (“**pga2021\_owgr**”) which merged **owgr2021** into **pga2021**.

**Step 2:** Remove “NaN” errors post-merge

1. As the OWGR CSV only included rankings up to number 998 in the world, “NaN” errors appeared in the OWGR column of **pga2021\_owgr** post-merge as some players who played on the PGA Tour in 2021 were outside the top 998 in the world rankings.
   * Used “fillna” to replace all “NaN” errors in the OWGR column with “Outside Top 998”.

**Step 3:** Add Ryder Cup teams to **pga2021\_owgr**

1. Manually created dictionary (“**rydercup\_dict**”) of all players who played on the Ryder Cup. Used dictionary as needed to include players and specify which team they played on.
2. Converted **rydercup\_dict** to a dataframe (“**rc2021**”).
3. Created new dataframe (“**pga2021\_all**”) which merged **rc2021** into **pga2021\_owgr**.

**Step 4:** Remove “NaN” errors

1. Not all players who played on the PGA Tour were eligible (e.g. players from Asia) or qualified to play in the Ryder Cup, meaning that “NaN” errors appeared in the “team” column of **pga2021\_all** post-merge (of **rc2021** into **pga2021\_owgr**).
   * Used “replace” function to replace “NaN” in the “team” column with “Did not compete”.

**Step 5:** Correct datatype of the SG columns in dataframe

1. The SG categories (e.g. SGP, SGT2G, etc.) which were imported from the 2021 PGA Tour Season Data CSV were stored as float64 due to a mixture of results being included (numbers and text – some text results of “NA”s were included in the raw data, primarily for events where SG data unavailable).
   * Created a new dataframe (“**pga2021\_final** “) and used “Dropna” on **pga2021\_all** to remove rows where “NA” was present in the “SGP” column (and therefore all SG columns).

**Step 6:** Create new PGA Tour winners dataframe (“**pga2021\_winners**”)

1. For the purposes of Strokes Gained analysis, data for just the winners of events on the PGA Tour in 2021 was required.
   * **pga2021\_winners** created using “loc”, where only 1’s appearing in the “Finish” column of **pga2021\_final** were maintained (1 representing the players final position in the event).

**Step 7:** Create new PGA Tour winners dataframe containing only SG data (“**pga2021\_winners\_sg**”)

1. As mentioned in section4 above, the 2021 PGA Tour Season Data CSV from ADA contained several columns relating to online fantasy sports games and were not relevant for this assessment.
   * New datadrame (**pga2021\_winners\_sg**) created using “drop” function on the using **pga2021\_winners** dataframe to remove these unnecessary columns.
2. “iterrows” was then used to create two new columns in **pga2021\_winners\_sg**:
   * SGP percentage – what percentage of SGT related specifically to putting for each players win; and
   * SGT2G percentage – what percentage of SGT related specifically to performance from tee-to-green (i.e. driving, approach to the green, etc.) for each players win.
3. The columns in **pga2021\_winners\_sg** were reordered to move the SG stats to the left-hand side of the dataframe.
4. **pga2021\_winners\_sg** was then indexed by player name and sorted by their OWGR.
5. The series **pga2021\_winners\_sg\_mean** was created to calculated the average of SGP and SGT2G percentages across all winners.

**Step 8:** Custom function:

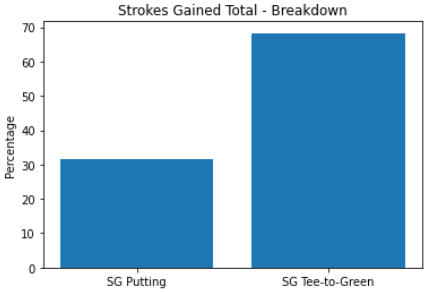
1. Created array (“**arr**”) of all players who won on the PGA Tour.
2. Custom function created which allows user to input any player and return a message telling them whether that player won on tour or not.

**Step 9:** Number of wins by player on PGA Tour

1. Used “groupby” on **pga2021\_winners\_sg** to count how number of wins each player had and sorted these in order of most wins to fewest.

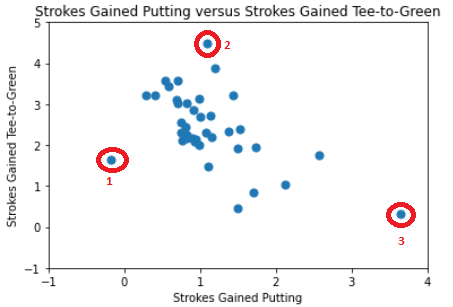
# Results

1. Breakdown of average SGT for all winners between SGP and SGT2G:



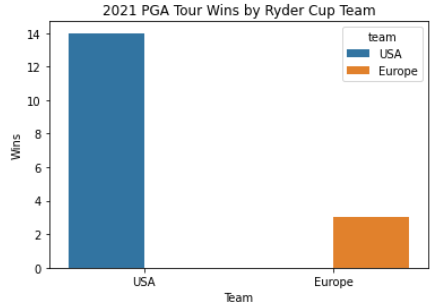
Figure

1. Winner's Strokes Gained Tee-to-Green (SGT2G) per round plotted against Strokes Gained Putting per round (SGP):



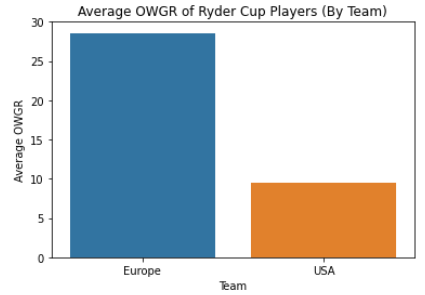
Figure

1. 2021 PGA Tour Wins by Ryder Cup Team:



Figure

1. Average OWGR of players on both Ryder Cup teams:



Figure

# Insights

1. **Breakdown of average SGT for all winners between SGP and SGT2G:**

Mark Broadie calculated that between 2004 and 2012 “*putting contributed 35%, on average, to victories on the PGA Tour, while off-green shots accounted for 65%*”. (Broadie, 2014)

As can be noted in Figure 1 above, performance tee-to-green (SGT2G) has become even more important as it contributed 69% to victories, with putting falling to 31% (frequently as low as 10%).

1. **Comparing SGT2G to SGP for all winners on PGA Tour in 2021:**

From Figure 2 above, players performed well in both SGT2G and SGP categories and both contributed significantly to their wins.

There were some outliers in the results however (see data points 1 to 3 in Figure 2):

1. There are some anomalies with Patrick Cantlay’s win in the 2021 Tour Championship:
   * Cantlay started the event on -10 (rather than even par) as he was the leader in the FedEx Cup.
   * John Rahm started on -6 and finished only one shot behind at the end of the event, due to by his SGP being much higher than Cantlay’s (similar SGT2G statistics). Cantlay therefore only won due to starting 4 shots ahead of John Rahm.
   * This was the only occasion that a player won on tour in 2021 with a negative SGP (mainly due to Cantlay starting on -10 as he would not have won the event if all players started on even par).
2. Byrson DeChambeau’s win in the 2020 US Open (delayed until the 2021 season due to COVID) was the highest SGT2G (an SGT) amongst the 2021 winners (both being over 0.5 shots higher than the next nearest result), contributing over 80% to his win.
3. Similar to DeChambeau’s performance noted above, Patrick Cantlay’s win in the BMW Championship is another outlier in the 2021 results, this time due to his SGP performance. This was the highest recorded SGP amongst tour winners in 2021 and contributed over 92% to Cantlay’s win.

Another anomaly with this result is that Cantlay won this event the week before he went on to win the Tour Championship (see point 1 above). He therefore followed the best SGP performance amongst the winners by the worst SGP performance (going from gaining 3.64 shots per round to losing 0.18 shots per round).

1. **How Ryder Cup teams compared before event started:**

Figure 3 shows total number of wins on the PGA by players on both teams. It is clear from this chart that players on the USA team were in much better form heading into the Ryder Cup as they accounted for over a third of all wins on tour during the season, with European players accounting for only 8% of wins.

Before the Ryder Cup, USA team were considerable favourites to win, despite their frequent poor performances in recent Ryder Cups. Figure 4 helps to back this up – the average world ranking of players on the USA team was 9 whereas the average on the European team was almost 30 (approximately 28.5).

USA went on to win the Ryder Cup comfortably and Figures 3 and 4 support why many saw the USA as outright favourites beforehand.

# References

1. Broadie, M. (2014). *Every Shot Counts*: Using the Revolutionary Strokes Gained Approach to Improve Your Golf Performance and Strategy. New York: Avery.
2. Dataset (Advanced Sports Analytics, 2021) *ASA All PGA Raw Data - Tourn Level.* [Downloaded 1 January 2022] (https://www.advancedsportsanalytics.com/pga-raw-data)
3. PGA Tour. (2016). *Strokes gained: How it works.* Available from: https://www.pgatour.com/news/2016/05/31/strokes-gained-defined.html [Accessed 1 January 2022]
4. PGA Tour. (2022). *Strokes Gained Explained:* Available from: https://www.pgatour.com/stats/stat.186.y2021.html [Accessed 1 January 2022]