Project Part 3

What is most important for success among PGA Tour golfers: driving, approach shots, or putting?

This is one of the most talked about aspects of golf today. With many golfers trying to increase their driving distance, I want to research if there is statistical evidence to support this transition. There are still golfers who don't hit the ball far at all and still compete with the big drivers, so I want to analyze any statistical trends in either direction. I will be able to measure success by analyzing how often players finish in the top 10 in tournaments and can use this as a basis of success (1). Earnings can also be used to analyze this as tournament prizing is based on resulting position and also the size/importance of the tournament itself. The best way to analyze PGA Tour success using the data I gathered, however, is to analyze how each aspect affects total strokes gained (2). Total strokes gained is a statistic that measures how well a golfer is playing relative to the rest of the players in the same tournament on the same day.

Data Description

Background Information

Golf is an incredibly difficult sport both mentally and physically. With the new analytics boom in all sports, golf has followed suit with consistently new technology in golf balls, clubs, and swing analysis. I want to research what aspects of the game truly make the difference between those golfers consistently competing for tournament titles and those fighting to make the cut in all of their tournaments. My goal is to find statistical trends among these golfers that can be useful to explain why certain golfers achieve more success than others on the PGA Tour. This data represents the population of golfers from the 2020-21 PGA Tour season. Because I want to analyze the modern game of golf, I say this is the entire population. The game is changing so much each year so I feel it is best not to generalize the

results of my research to the entire PGA Tour history. I used both PGA Tour golf data and data from ESPN. ESPN provided general golf statistics and earnings (1), while the PGA data was more advanced golf statistics (2) that I specifically wanted to analyze further. Both sites simply collected the data by doing their own tracking of PGA Tour events throughout the 2020-21 season and presenting the statistics in their own way. The data I gathered will give me the opportunity to go into great depth to further analyze my research question and to give the reasoning behind my eventual answer.

Data Collection

I collected the PGA Tour data by copying the statistics from both the PGA Tour (2) and ESPN (1) websites into an excel spreadsheet. From there, I was able to do some basic manipulations to make sure each column was formatted correctly. After that, I saved the spreadsheet as a csv to my computer to be able to read the data into R most efficiently.

Data Manipulation

I made some data manipulations. First, I examined all of the statistics I gathered from both data sources and removed the ones that I felt were of no importance to answering the research question. I wanted to merge the datasets by golfer name, so the next thing I did was change the golfer names for the datasets I acquired from the PGA Tour website (2) from "PLAYER.NAME" to just "NAME" as they were listed in the ESPN dataset (1). The final thing I did was create a new categorical variable called "AnyWins" that had the string value of "Win(s)" if that golfer had at least one win over the course of the season and a value of "NoWins" if they didn't.

About the Data

Each row in the dataset is an individual player who participated in the 2020-21 PGA Tour season. The variables in the dataset for each player are as follows:
* NAME: Name of PGA Tour golfer
* RK: Final ranking of golfer
* AGE: Age of golfer
* EARNINGS: Earnings of golfer
* RNDS: Number of rounds each golfer played
* CUTS: Number of cuts made by each golfer

* TOP10: Number of top 10 finishes by each golfer
* WINS: Number of wins by each golfer
* SCORE: Average score per round by each golfer
* DDIS: Average driving distance on Par 4s and Par 5s
* DACC: Driving accuracy on Par 4s and Par 5s, percentage of fairways hit
* GIR: Percentage of greens hit in regulation by each golfer
* PUTTS: Average number of putts per hole for each golfer
* AVG.CHS: Average club head speed of drives for the golfer
* AVG.BS: Average ball speed of drives for golfer
* AVERAGE: Average number of strokes gained: the number of strokes per round the player was better or worse than the field average on the same course and event
* TOTAL.SG.T: Average number of strokes gained off the tee each round
* TOTAL.SG.T2G: Average number of strokes gained after tee shots and before putting each round
* TOTAL.SG.P: Average number of strokes gained putting each round

Potential Issues

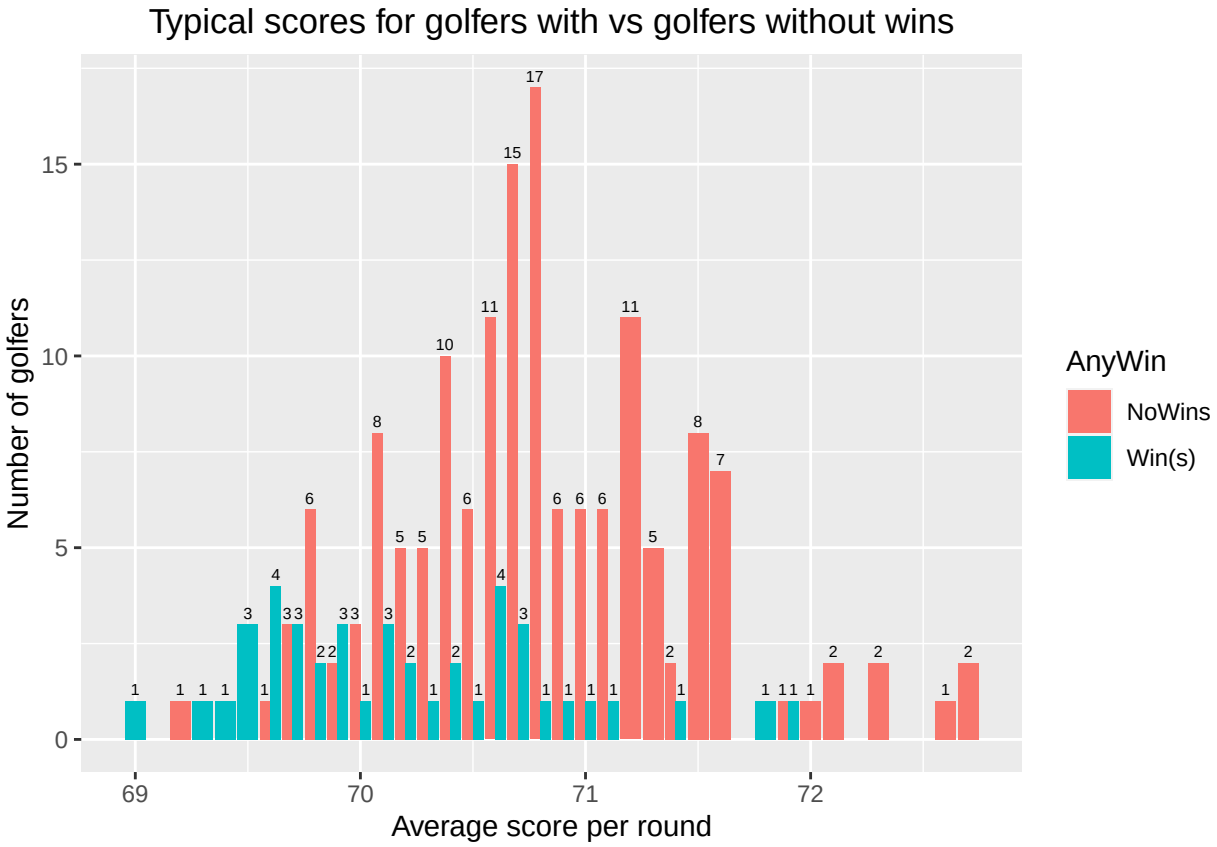
When comparing the data between both sites, I realized that there was a slight difference in the number of rounds for each golfer in their 2020-21 seasons. The number of rounds was slightly higher from the PGA Tour website (2) and I believe this is due to the fact that they included their postseason tournament in their final statistics, while ESPN (1) did not. The numbers aren't significantly different, but because of this, I will not be comparing any 'total' statistics. Regardless, there is enough data for each golfer that the 'average' statistics can be trusted and appropriately analyzed without having to worry about too small of a sample size helping or hurting any golfer.

Data Summaries

Numerical Summary

```
big_pga %>%
  ggplot(aes(x = SCORE, fill = AnyWin)) +
  geom_bar(stat = 'count', position = 'dodge') +
  geom_text(stat = 'count', aes(label = ..count..),
```

```
          position = position_dodge(.1), vjust = -0.5, size = 2) +
labs(title = "Typical scores for golfers with vs golfers without wins",
     x = "Average score per round",
     y = "Number of golfers") +
theme(plot.title = element_text(hjust = 0.5))
```



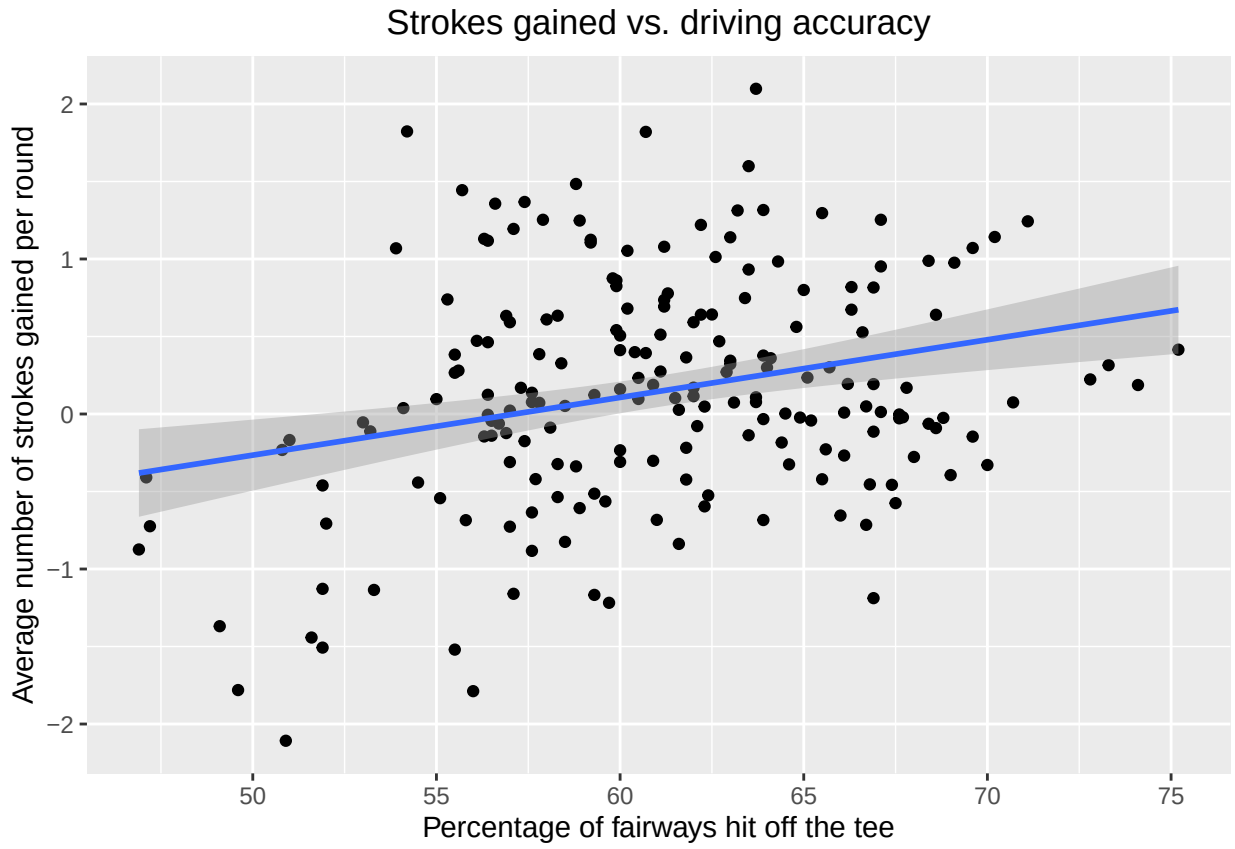
This is a perfect graph for showing what the typical scores are of golfers that won a tournament this season and the typical scores of the golfers who didn't. It is from this numerical summary that the golfers who had at least one win on the PGA Tour during the 2020-21 season tended to average a lower score over every round they played during the season. It is also clear that only a small amount of golfers on tour actually win a tournament. This shows that winning on tour isn't a fluke and the most consistent golfers win the most tournaments. This is a good baseline when first looking at the data to understand that there is a consistent trend that can be seen with PGA Tour winners.

Graphical Summaries

```
big_pga %>%
```

```
  ggplot(aes(x = DACC, y = AVERAGE)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  labs(title = "Strokes gained vs. driving accuracy",
       x = "Percentage of fairways hit off the tee",
       y = "Average number of strokes gained per round") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `geom_smooth()` using formula 'y ~ x'
```

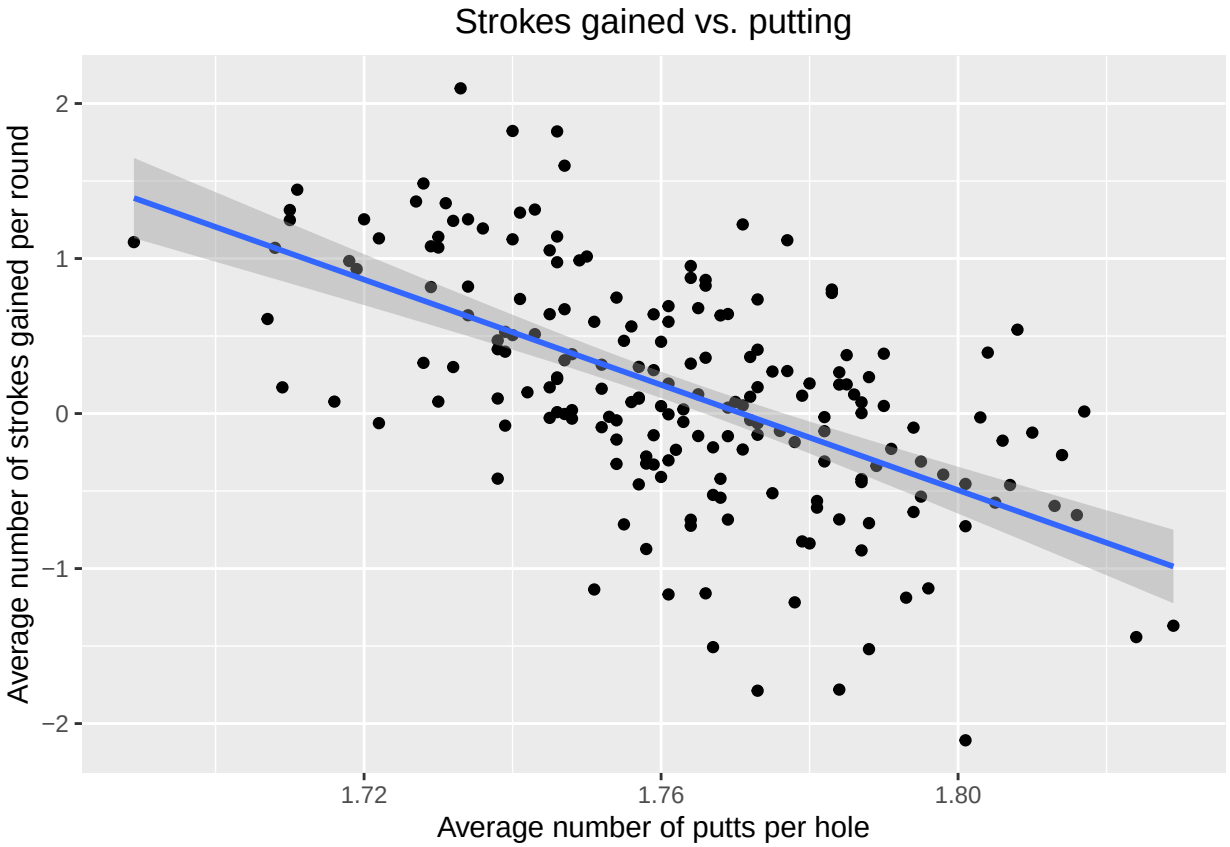



```
big_pga %>%
  ggplot(aes(x = GIR, y = AVERAGE)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  labs(title = "Strokes gained vs. greens in regulation",
       x = "Percentage of greens hit in regulation",
       y = "Average number of strokes gained per round") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `geom_smooth()` using formula 'y ~ x'
```


```
big_pga %>%
  ggplot(aes(x = PUTTS, y = AVERAGE)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  labs(title = "Strokes gained vs. putting",
       x = "Average number of putts per hole",
       y = "Average number of strokes gained per round") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



These three plots were important to include in this report because they help me begin to have an understanding about the answer to my research question. My question was asking whether driving, approach shots, or putting was the most important factor of success among PGA Tour golfers, so I plotted each against average strokes gained per round by golfers. Average strokes gained is the best measure of success because it not only analyzes how well a golfer plays, but it compares it to the rest of the golfers in a particular tournament. The graphical summaries above give an indication that greens in regulation percentage (approach shots) and average number of putts per hole have a stronger correlation with strokes gained per round than driving accuracy. This indicates that approach shots and putting may be more important than driving in terms of success on the PGA Tour. I now have an idea of what to research further when I analyze the data more in the future.

References

1. https://www.espn.com/golf/stats/player/_/season/2021
2. https://www.pgatour.com/stats.html