Intro

I suck at golf. According to the USGA, I am currently a 20.6 handicap. This means that even on my best days, I still average worse than a bogey on a given hole. My goal is to become a **15 handicap**. Better yet, I'd prefer to reach that milestone before my dad, who also sucks but does manage to practice more than me.

I would love to get better, but that takes lots of time and effort. Given I work full-time during the day, take classes in the evening, like to travel, and do all these other random things, it is *not realistic* to think that I could somehow quickly improve my golf game by 5 strokes.

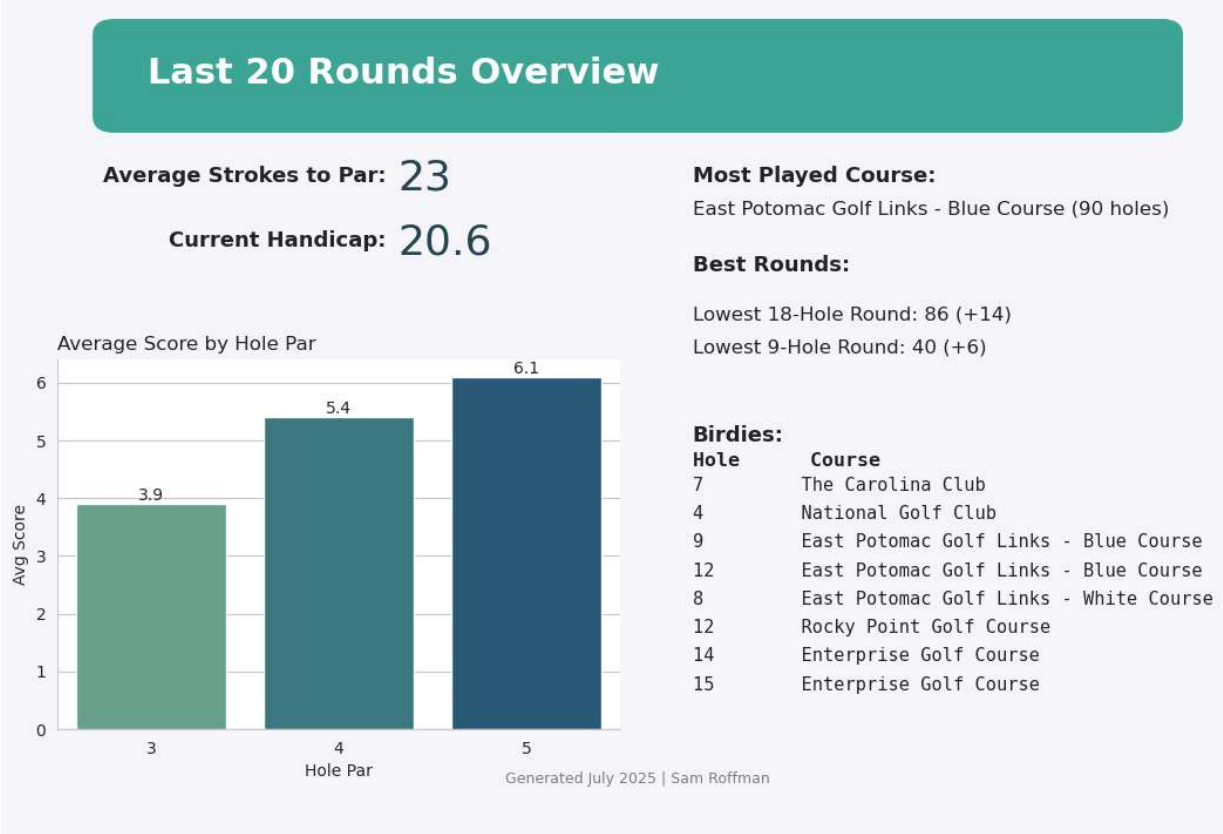
The *golf swing* is only part of the game. Course management and understanding strategy is a major part of scoring, and getting this is often how high handicaps see improvement. I thought: What if I can use data about my game to get a better understanding of where I gain strokes, or more likely lose strokes, and use my data science background to come up with a gameplan to get me to a 15?

I track all my rounds on 18 Birdies, which allows me to keep stats on my **fairways**, **greens**, **chips**, and **putts** per round. It also has each hole's **handicap rating**, allowing me to see how I fare on the easiest and hardest holes on the course.

Using this data, and a little bit more that I'll get to, I set out to better understand my golf game, and hopefully come away with a gameplan to get me to a **15 handicap**.

Overview

To begin, let's level-set with seeing where I've been over my last 20 rounds. Below is an overview of all the rounds making up my handicap, my average scores by different pars, best rounds, and, just for fun, where I've made my few birdies.



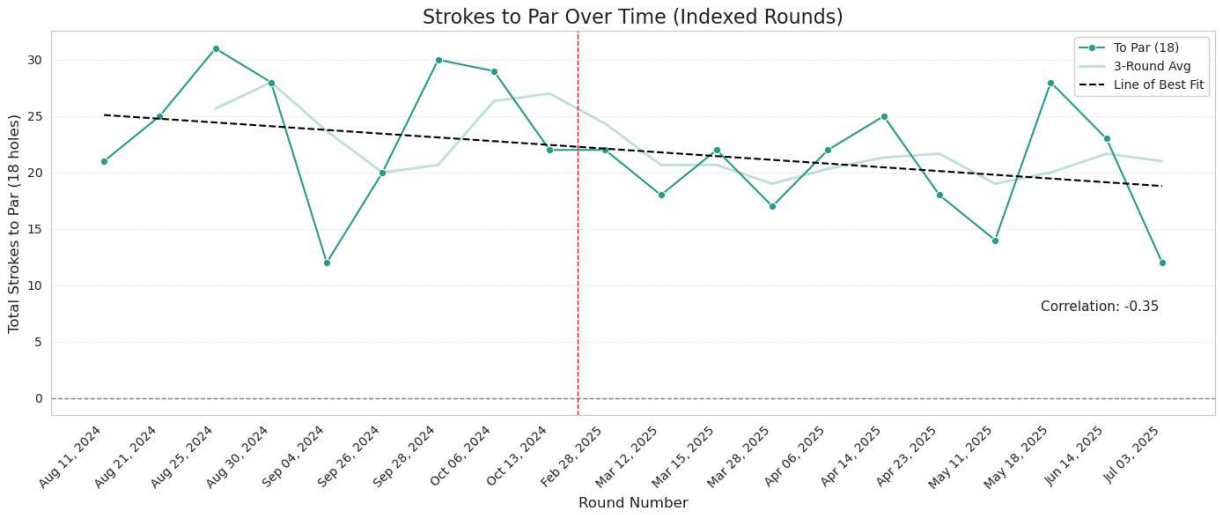
Exploratory Data Analysis (EDA)

Ultimately, I would like to build a statistical model, using machine learning to understand the key components of my score. Before that, however, it is important to understand the data we are using. In data science, this is done with *Exploratory Data Analysis*, or EDA. Below, we will look at my recent trends over time, followed by some breakdowns of scores and other stats within different groupings of holes.

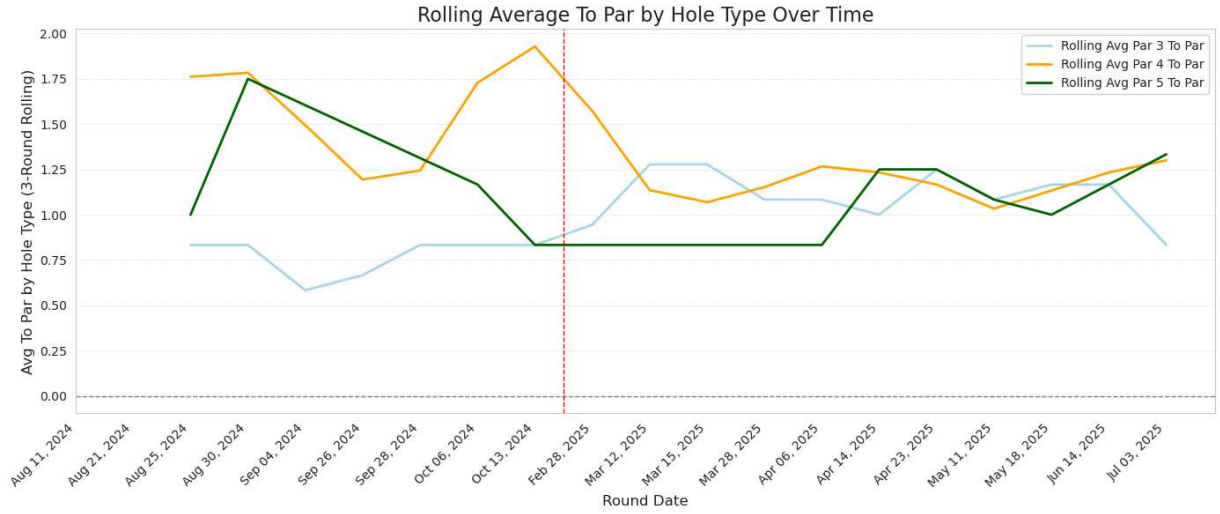
Once we have a good subjective understanding of the inner workings of my scores, we can move to machine learning or more complex statistical analysis to attempt to gain an **objective, mathematical answer** to the question: *What course management strategy can I follow to become a 15 handicap?*

Trends

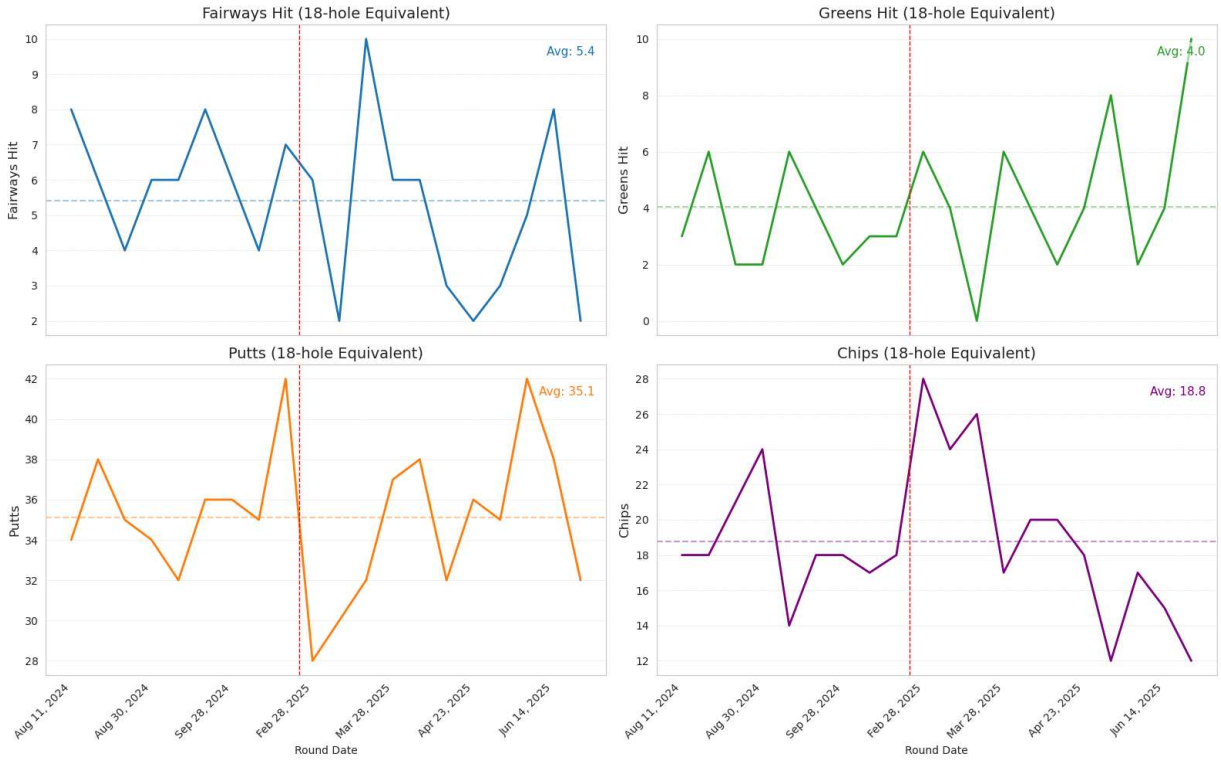
Below, I've plotted my total score over time, accompanied by a 3-round rolling average, and a trendline.



Our trendline shows a `correlation coeficient` of **-0.35**. This indicates we are, infact, **Trending!** We have statistical evidence that there is a medium-strength, negative relationship between round number (least recent to most recent) and score. This means that from my last 20 rounds, as I have played more, my expected score has gone down. (Approx. 1/3 of a stroke per round).



Here, we are looking at my average scores to par by par 3s, 4s, and 5s. While there aren't any significant downward trends here, I am pleased to see an apparent decrease in variance between the three lines. This indicates less volitility in my recent rounds compared to my older ones, which is always a good thing.



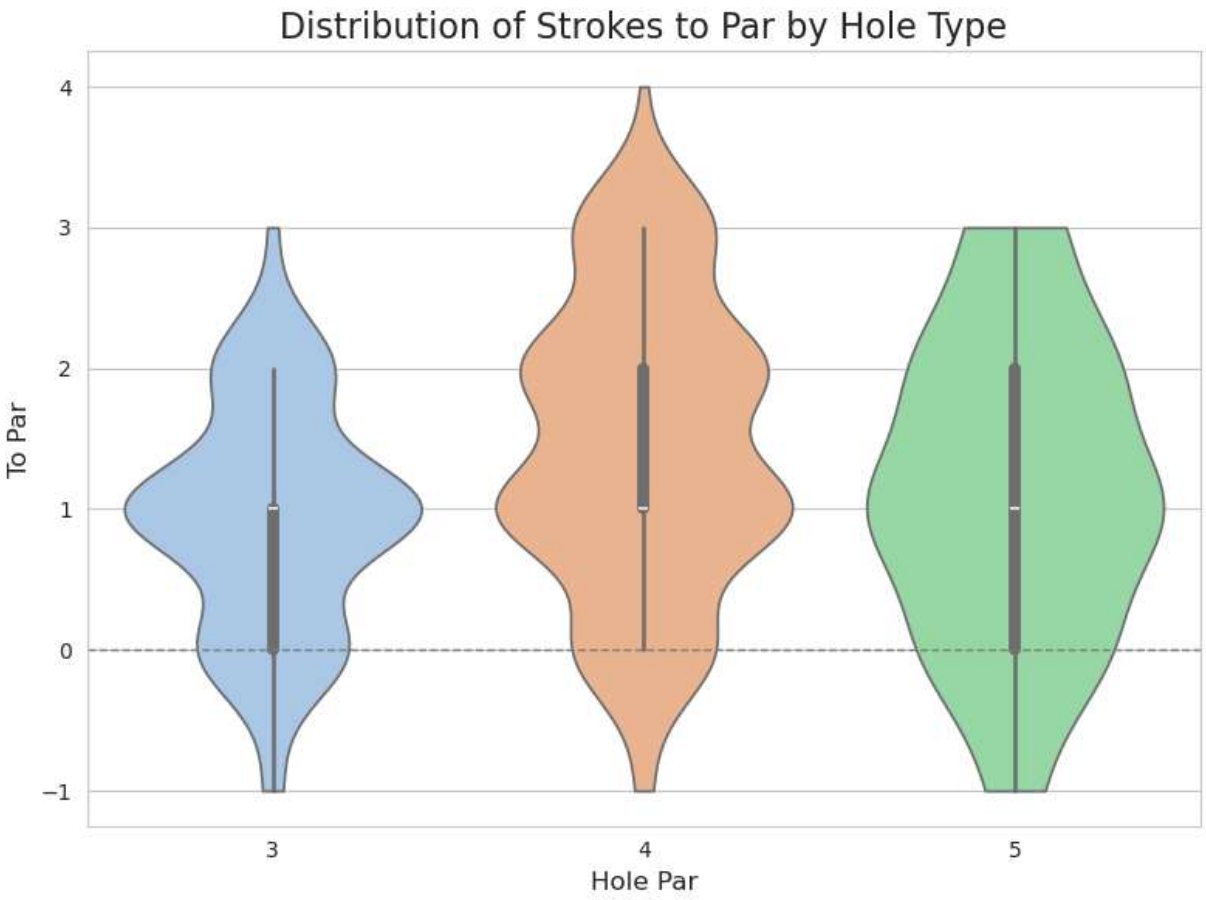
Finally, let's look at my stats from this time period. My fairways and putts are all over the map, but there does appear to be good recent trends in greens and chips. This makes sense, as the more greens I hit, the less I have to chip.

Probably most concerning are my 20-round averages in chips and putts. 2 putts per hole isn't bad, but coupled with over a chip per hole is not a good sign. While 1 chip and 2 putts usually equals a bogey (good for me!), the volatility in my game often means that I will end up with scores in the mid to high 90s, even with these stats.

Hole-by-Hole Breakdowns

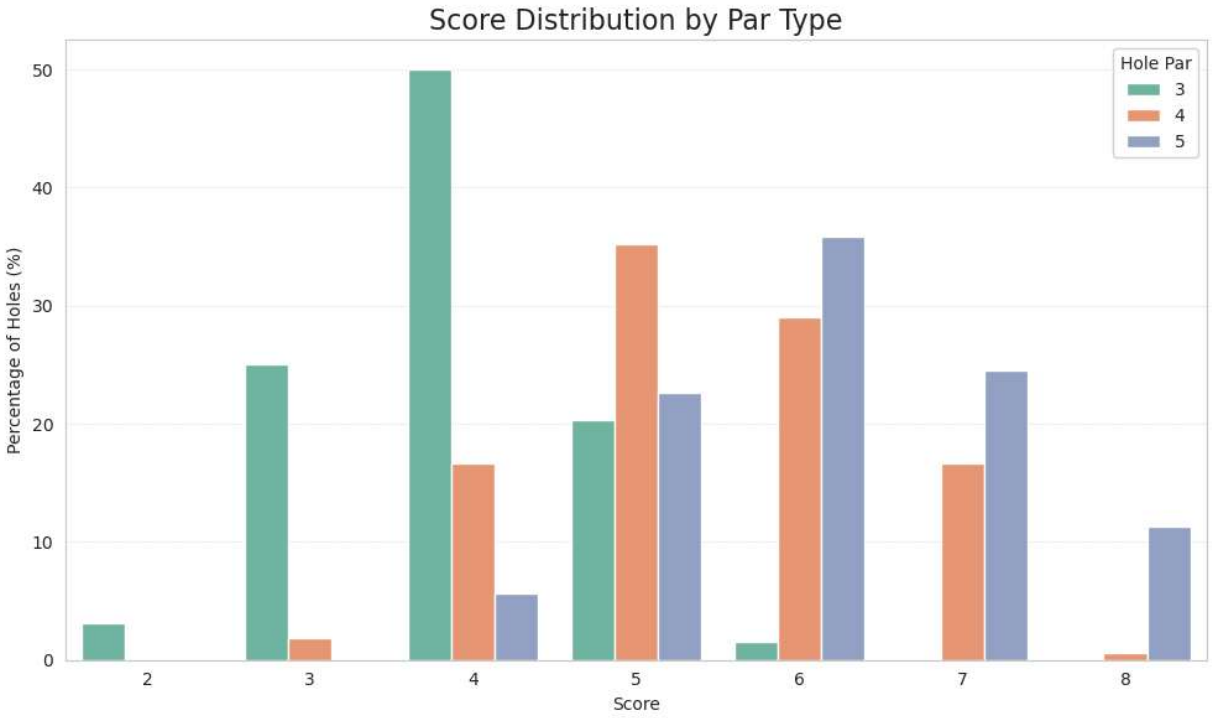
Next up in EDA, I want to do some `groupby` functions to see how the data looks in different subsets. This will help us look for key differences in scoring or stats across different scenarios that occur throughout a round. Below, we will look at different **pars**, **handicaps**, and **hole numbers during a round**.

Breakdown by Par



This is a violin plot, which is one of my favorie plots for comparing distributions across groups. Think of it as a more detailed boxplot, where the highest values are on top, the middle values are in the middle, and the lowest ones are at the bottom. Violin plots are wider where there is a greater concentration of data. For example, I make lots of bogeys on par 3s, so for the plot on the left, the violin is wider around the value Y = 1 (1 stroke over par).

In this plot, it is clear there is a lot of variance in my scores on par 4s and 5s, since the violins are wider at the extreme values. I tend to do more consistently on par 3s, which is great.

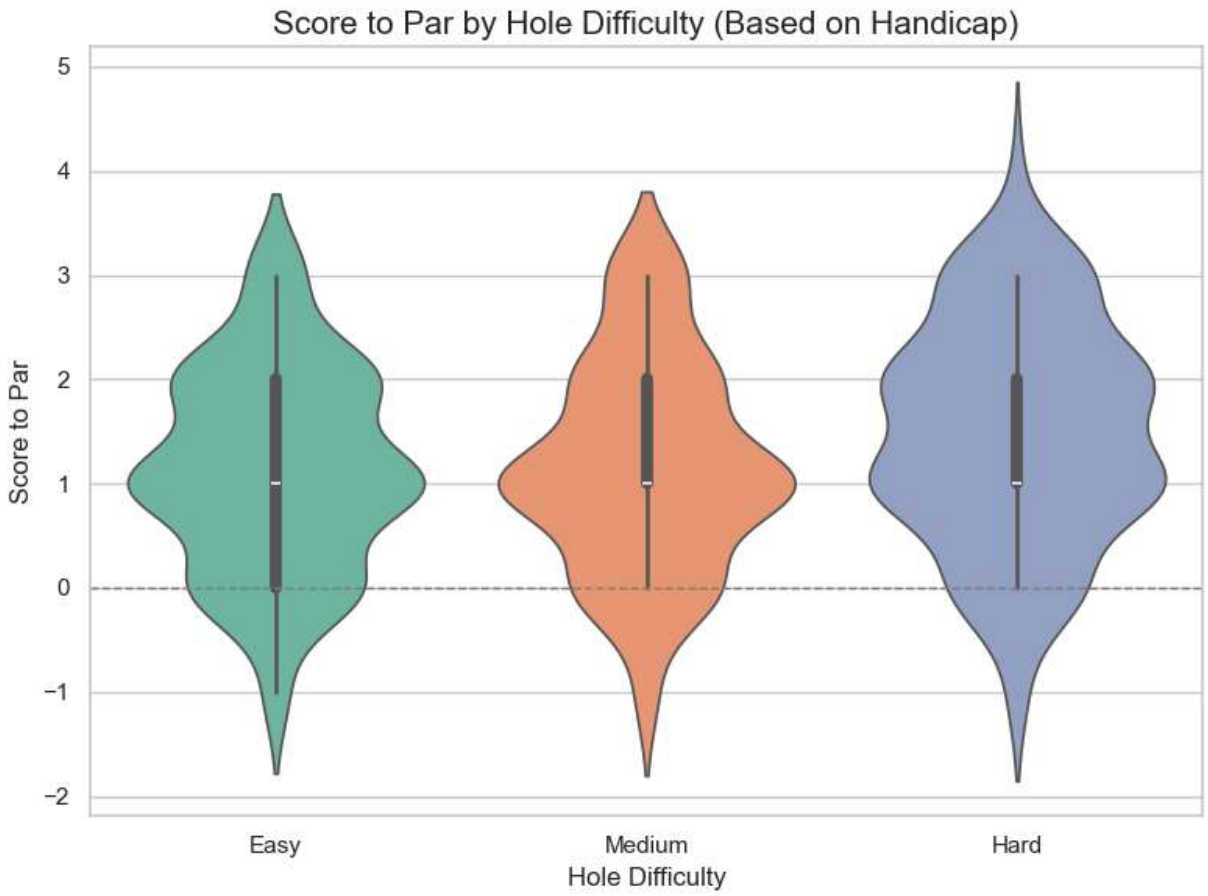


Here is another way of looking at the data from the violin plot. You can see I make 4 on about 50% of my par 3s, and actually make 3 more often than I make 5 or 6. This is again great news.

More concerning is my record on par 4s. I make double nearly 30% of the time on par 4s, and triple over 15% of the time. Comparing this to par 5s, those values are 25% and 11% respectively. Limiting blow ups on par 4s is going to be key moving forward.

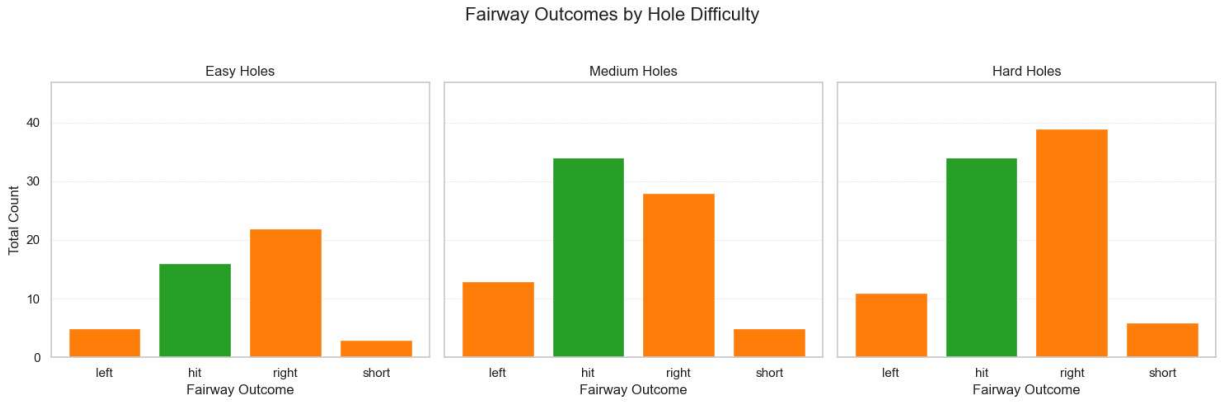
Breakdown by Difficulty (Handicap)

Next, lets look at holes of different diffculties. For ease of analysis, I have used the handicap system to define holes as **Easy** (handicaps 13-18), **Medium** (handicaps 7-12) and **Hard** (handicaps 1-6). Obviously, these ratings are somewhat subjective, and a *medium* hole on a hard course may be harder than a *hard* hole on an easier course. Nonetheless, this is the system we have.

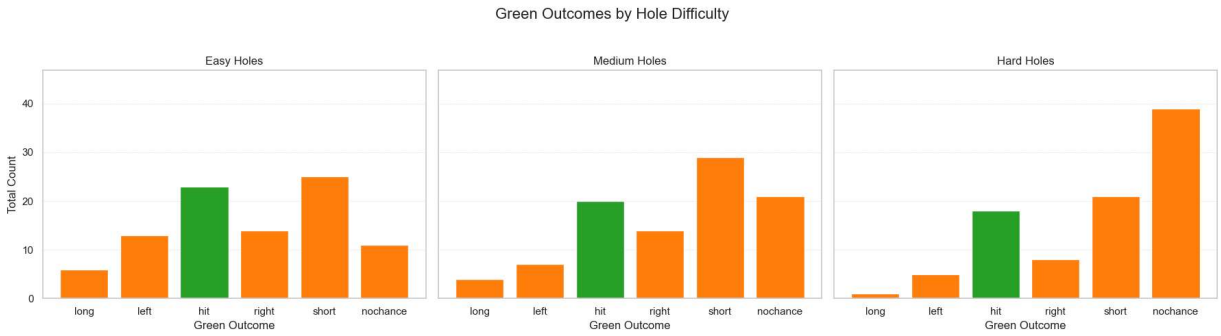


Looking again at the violin plot, we see a trend that generally makes sense. The center (mean and median) of the distribution of my scores rises as holes get harder.

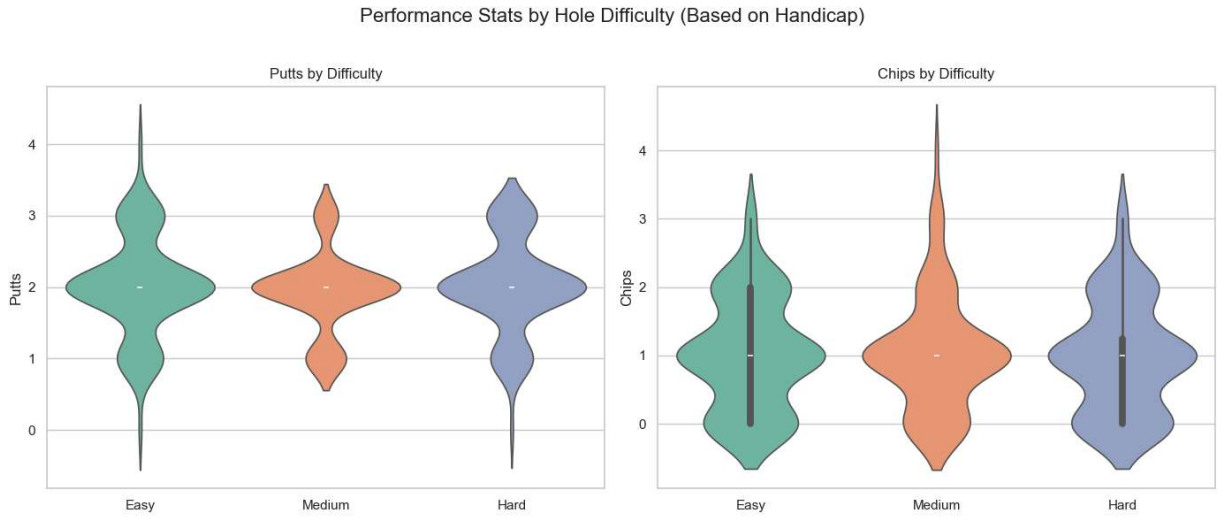
One interesting note: There seems to be a higher concentration of doubles (Y=2) on easy holes than hard holes. This may be due to the fact that many easy holes are *Risk / Reward holes*. Maybe I go for the green in 2 on an easy par 5 - if I make it, I have an easy birdie or par. If not, I may be staring at double or worse.



These barplots show the distribution of my fairways hit and missed. There isn't much to pull out of this, but it is interesting to see that I hit more fairways on *medium* holes than *easy* ones. The raw counts are higher because easy holes are often par 3s, but the `hit` bar is higher than all the rest for *medium* holes.



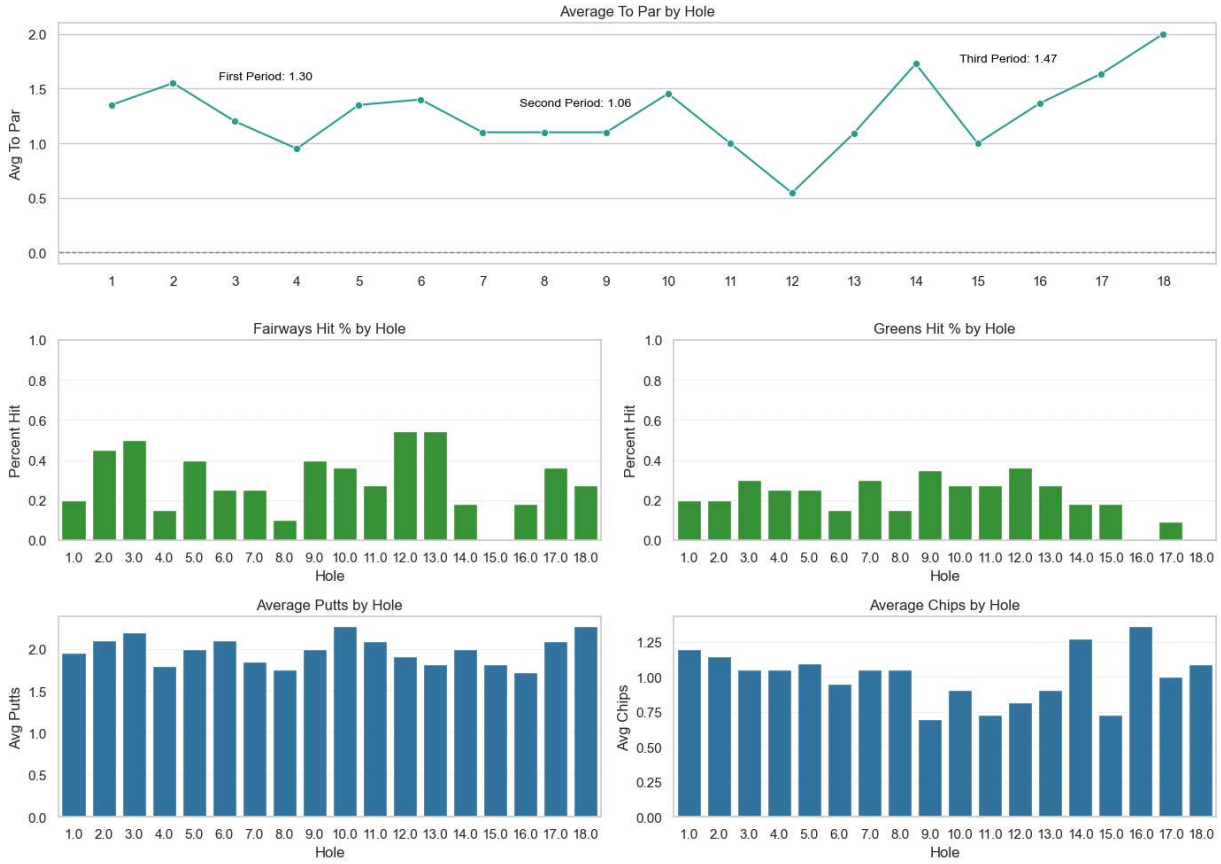
Now we are starting to see something interesting. *Easy* and *Medium* holes have fairly similar green dispersion pattern, but on *Hard* holes, I miss the green more, and have `No Chance` to hit the green almost **40 %** of the time. This is less a comment on my iron play, and more due to the fact that I am not getting off the tee well on hard holes. It doesn't matter if I miss left or right, having no shot at the green (or taking a penalty) is not good for lowering scores.



Just a couple more violin plots, since they are of course my favorite. Not much to see here. I sort of thought there would be way more putts and chips on hard holes. There are slightly more 3 putts, but for the most part the distributions look pretty similar across the groups

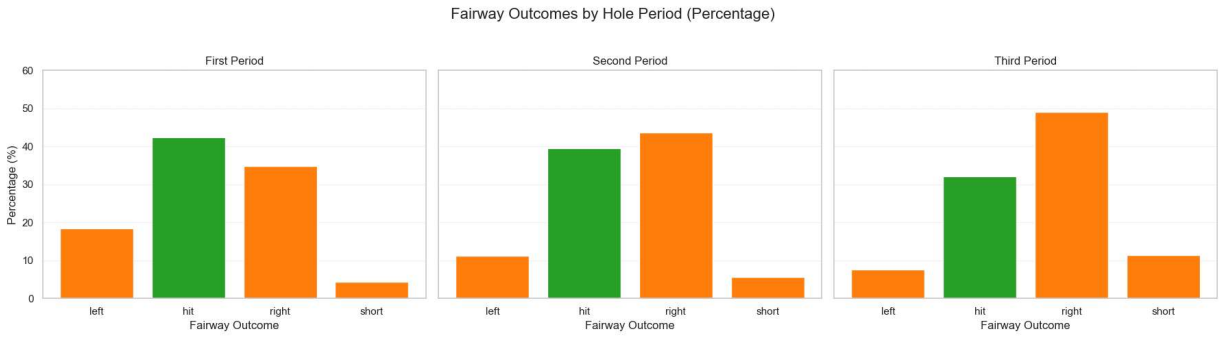
Breakdown by Hole Number

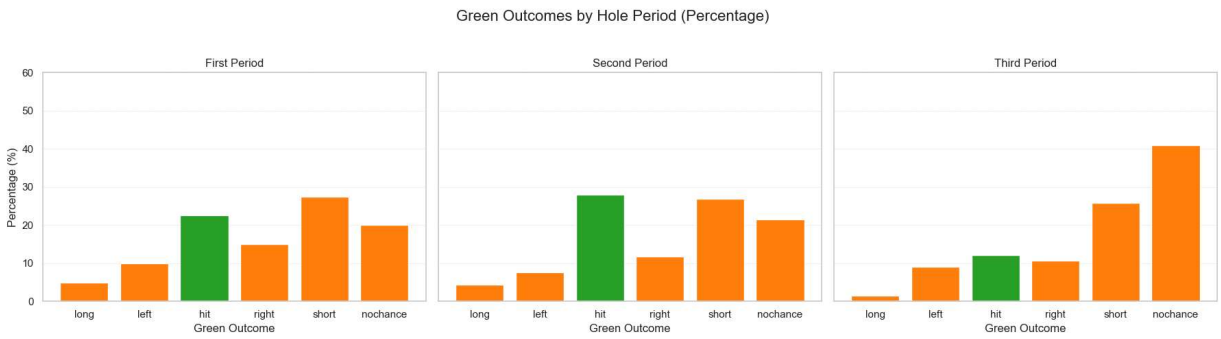
Based on emperical evidence, I think I have a tendency to flame out at the end of the round. Whether it is fatigue or nerves about potentially breaking a scoring barrier and reaching a personal Megabonus, I find I often blow up near the end. Let's find out if that is backed by the data, breaking the holes into three `Periods` (holes 1-6, 7-12, and 13-18).



The answer to my question is yes. The numbers do back up my theory that I flame out at the end. In the `Third Period`, I average almost 1.5 strokes over par, compared to 1.18 for the first two periods.

The more important ask is what causes these higher scores. While there are some spikes, it doesn't look like I get particularly worse and chipping or putting late in the round. The bigger drop off is in fairways and greens. I need to get off the tee better, to set myself up for a better 2nd and 3rd shot.





This breakdown backs up what we were seeing before. The majority of my missed greens in the Third Period are because of bad tee balls, leaving me no chance to hit the green. In fact, the proportion of holes like this is almost identical to the 6 hardest holes by handicap (~40%). I am getting off the tee equally poorly in the hardest holes and latest holes on the golf course.

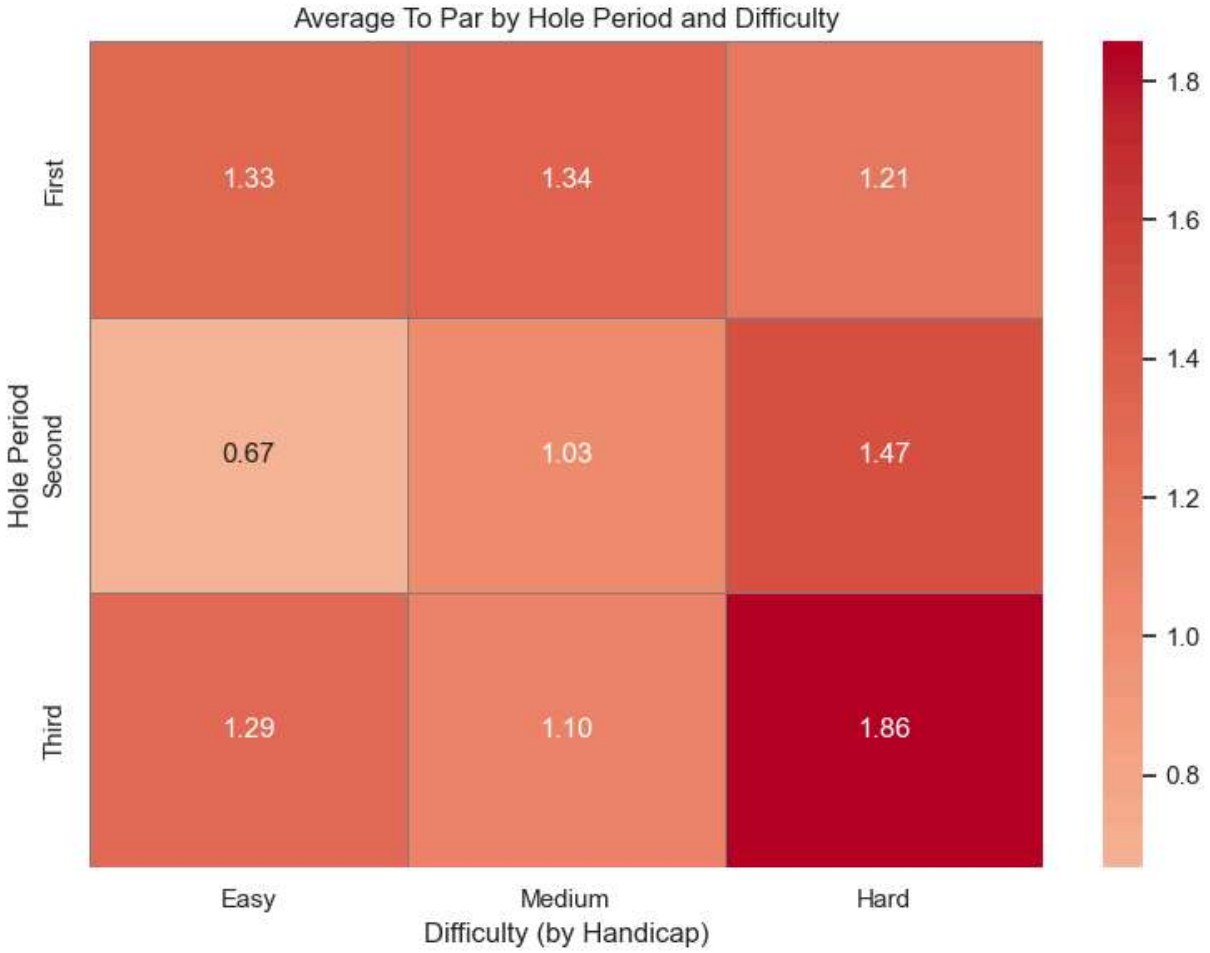
Now we have a better understanding of the data. Let's dig a little deeper with some `Conditional Expectations`

My Gameplan - Conditional Expectations

To build a gameplan for myself, I first need to understand how the `Expected Value` (EV) of my score changes in different situations.

We will look at different combinations of hole types, as well as what happens once I hit my tee shot in different locations.

Hole Period and Difficuly -- Expected Values



Above, we have a table of my expected score to par in different combinations of hole difficulty and time throughout a round. For example, on hard holes late in the round, I average 1.86 strokes to par. On easy holes in the middle of the round, I average 0.67.

Below, we will look at my average strokes to par, given different tee balls, and given different approach shots.

Out[169…
```
fairway
hit      1.035714
right    1.505618
left     1.586207
short    2.071429
Name: to_par, dtype: float64
```

Out[175…
```
green
hit         0.196721
long        1.181818
left        1.360000
right       1.361111
short       1.386667
nochance    1.971831
Name: to_par, dtype: float64
```

Everything we are seeing to this point makes sense. "Hit fairways & greens = Good. Miss short of the fairway = Bad."

Now, let's combine these variables to dig a little deeper.

Out[312...

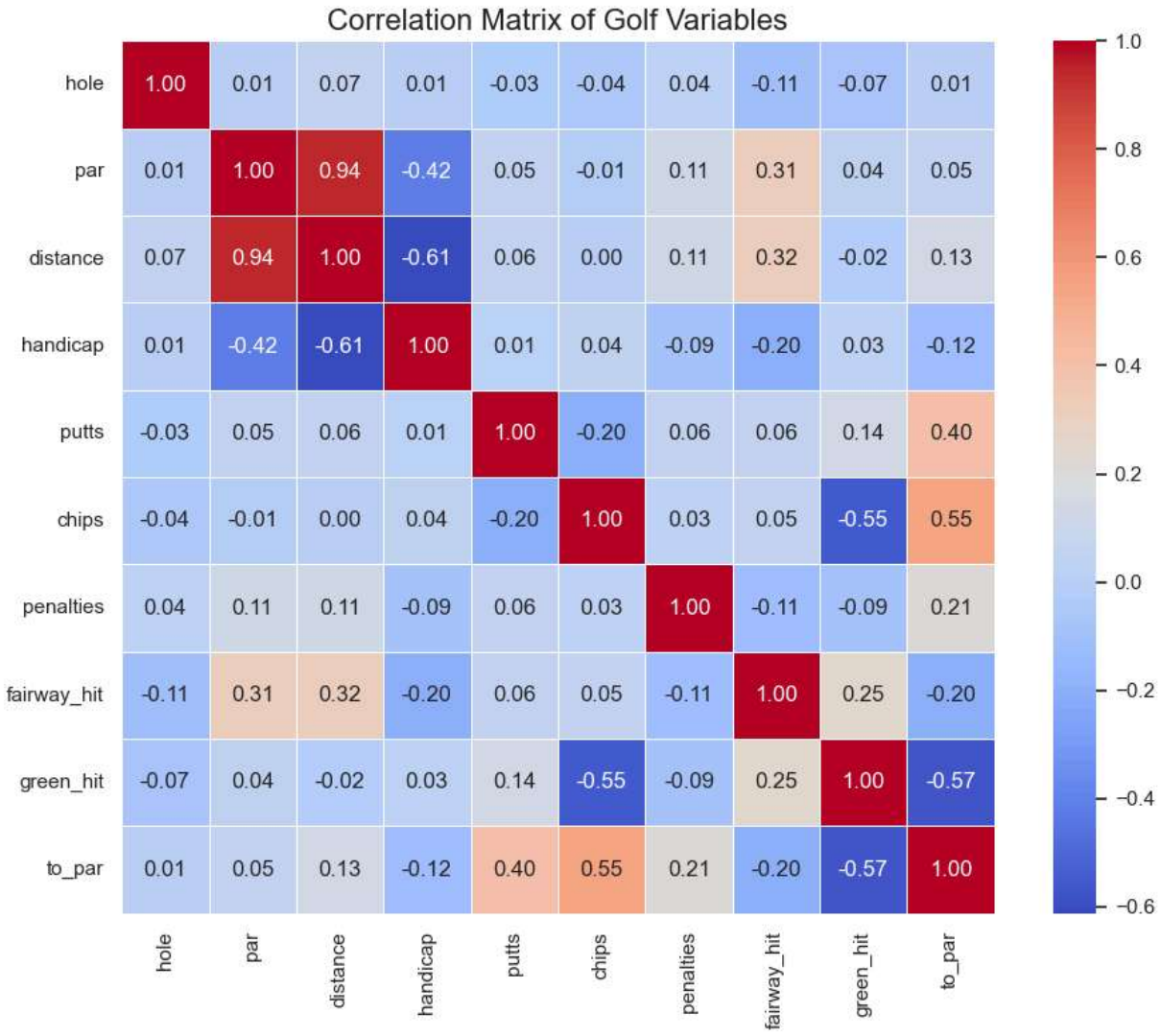
period	First			Second			Third		
difficulty	Hard	Medium	Easy	Hard	Medium	Easy	Hard	Medium	Easy
fairway									
hit	0.80	1.22	1.25	1.00	0.00	0.40	1.67	0.80	nan
left	1.40	1.20	1.50	0.50	2.00	1.00	nan	2.50	nan
right	1.71	1.33	1.83	2.25	0.50	1.00	1.80	1.40	1.00
short	1.00	3.00	3.00	nan	2.00	1.00	2.25	nan	nan

The pivot table above shows how different fairway misses affect my score on different holes. Interestingly, left misses seem to hurt less than right misses, likely because a right miss can often come with a big slice. It also looks like this slice is worse at the beginning of the round. The right miss hurts me a lot more early than it does late. Keeping a smooth tempo early in the round until I get comfortable could help me mitigate large numbers early on.

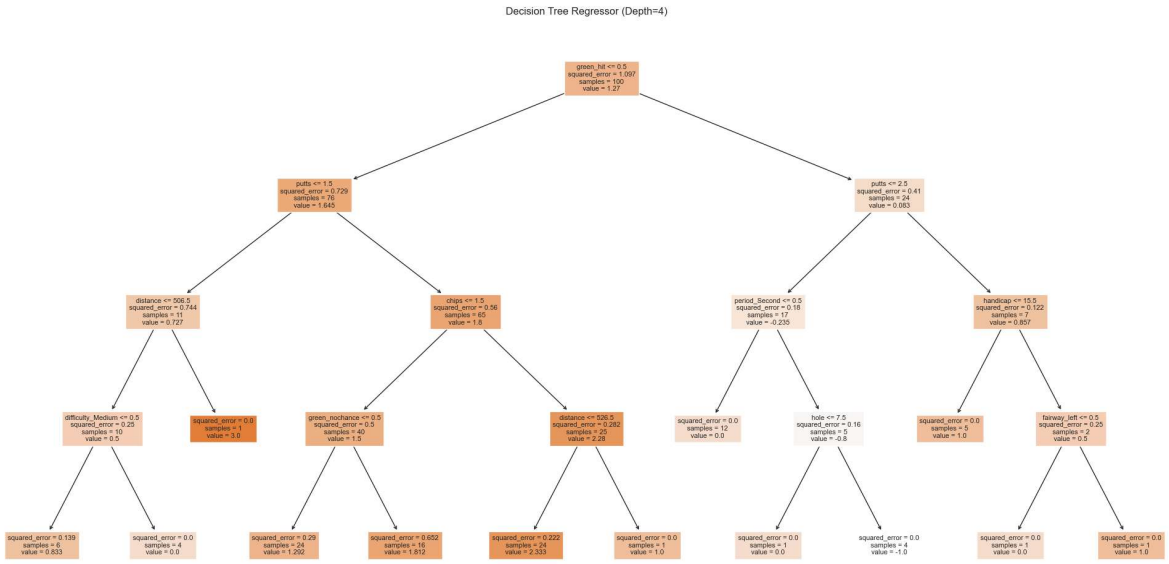
Building a Model

All those analytics are great, but it is time to build a machine learning model to see how the computer synthesises all of this information. Given I do not have a lot of data, and I am dealing with categorical, nonlinear data, I have settled on using a `Decision Tree Regressor`. I like this option particularly because it is not a black box - we can visualize the tree and interpret how it makes its decisions predicting my score.

The important part of this exercise is not predicting scores, but understanding the impacts of different features in creating the predictions.



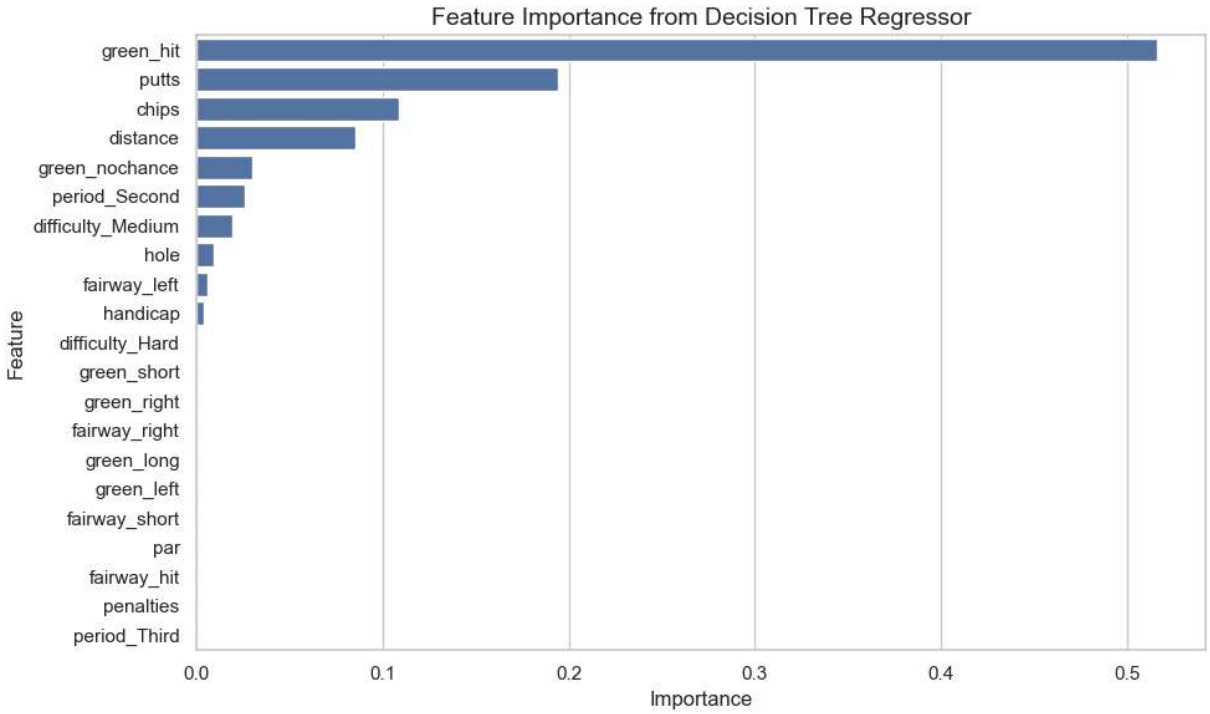
Above is a correlation matrix. This shows how much each numerical variable is correlated with each other. Darker colors indicate more correlation, and hue indicates directionality. According to this, hitting greens and minimizing chips and putts are the most important variables to lower my score `to_par`.




```
R-Squared: 0.41118309620596205
```

As described, here is our Decision Tree visualized. This model has an R-Squared value of 0.411, meaning it explains about 40% of the variance in my score - pretty solid for such a random value, with such little data.

You can read the tree yourself. Each node has a decision at the top of it. For example, the first decision point is `green_hit <= 0.5`. In english, this is asking: "Did I miss the green?". Under the node, there will be two arrows. The arrow to the left answers "Yes" to this deision point, and the arrow to the right answers "No". You can see that in the node to the right under this first decision point, the value is `0.083`. This means that if I hit the green, this is my expected score to par. You can continue following the logic of the tree, understanding that each subsequent decision point includes the answers to all previous decision points as its base scenario. At the bottom, the terminal nodes represent the expected scores to par for each of the 13 identified most influential combinations of predictors and values.

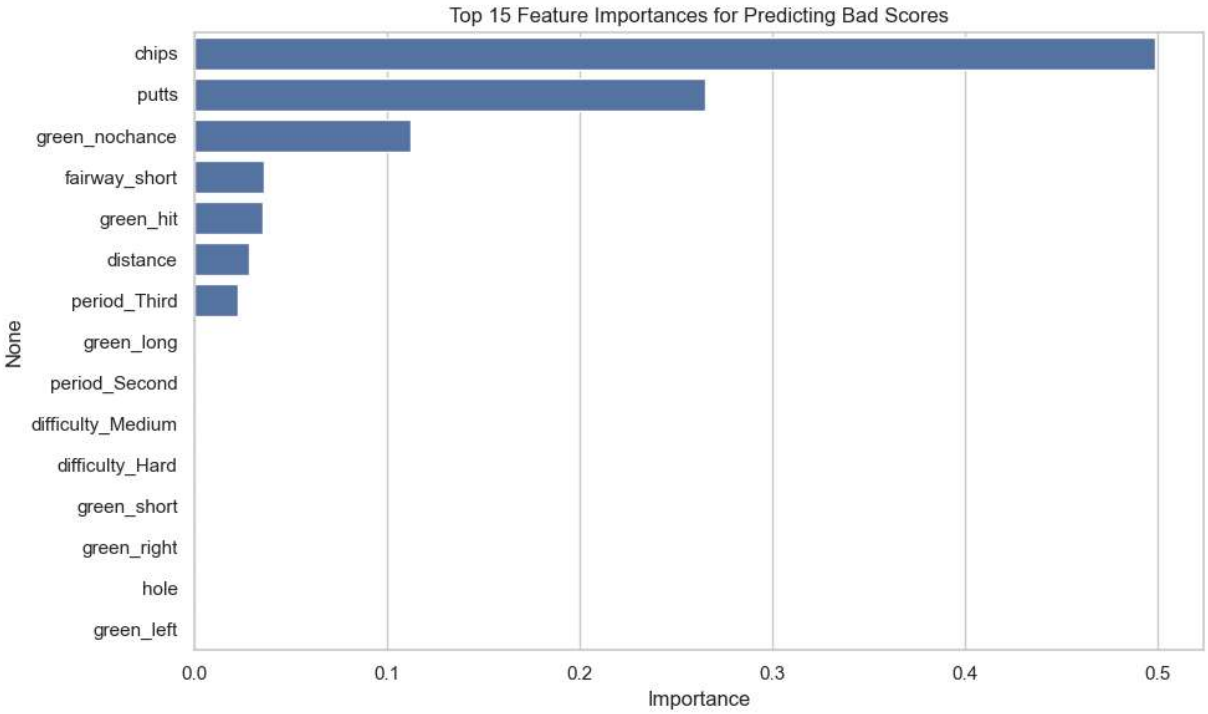


This is a feature importance plot, which tells us exactly what it sounds like. Which predictors are most important for predicting score? For this model, the answer is:

- Hitting greens
- Putting
- Chipping
- Hole distance

Doubles or Worse

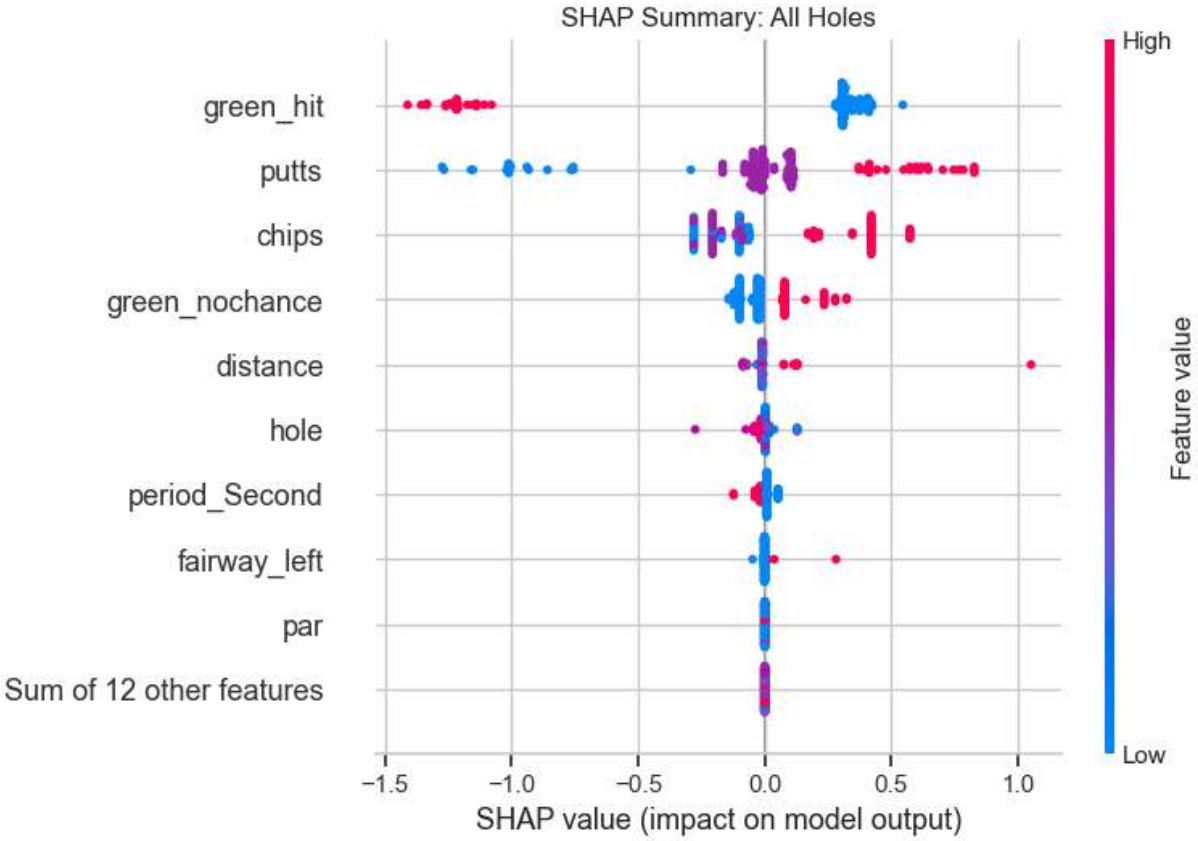
Next, I wanted to adjust the model a little bit. Predicting scores is great, but as a high handicap I really need to focus more on eliminating the blow-ups. This model below is a `Decision Tree Classifier`, attempting to predict which holes will score a double or worse.

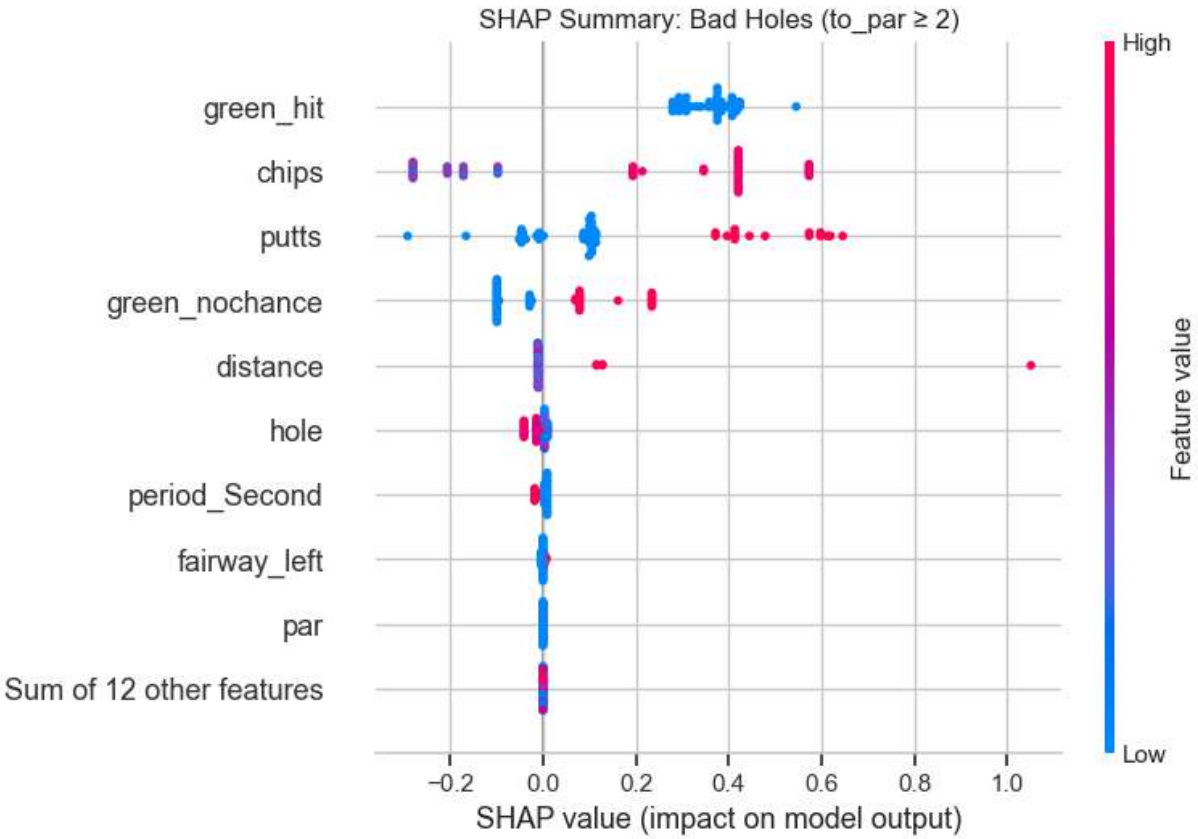



```
              precision    recall  f1-score     support
0              0.928571  0.902778  0.915493   72.000000
1              0.875000  0.907407  0.890909   54.000000
accuracy       0.904762  0.904762  0.904762    0.904762
macro avg      0.901786  0.905093  0.903201  126.000000
weighted avg   0.905612  0.904762  0.904957  126.000000
[[65  7]
 [ 5 49]]
```

This model peformed very well, with all accuracy metrics being above 0.85. We should take this with a grain of salt, as I did not split up training and testing data. There is not enough data here to do that and get an accurate ML model. That doesn't really matter, however, as the most important takeaways here come from Feature Importance.

We can see the big difference clearly. The most influential feature for predicting bad holes is `chips`. This is different from the first model, and important to remember on the course: The model tells me that my tee shots and approaches are good enough to not be a huge predictor of when I blow up. My blow ups are primarily caused by my short game. This means that even when I go OB, I need to remember that my most important shots are yet to come.





`Shapley Values` show feature importance on a case-by-case basis. While feature importance plots show which predictors are important to the model in general, Shapley values tell us: For a given prediction, which specific values of each predictors were most

influential? For example, on a hole where I 4 putt, putts are going to be way more influential than a hole where I two-putt, but hit my tee shot OB.

These plots (first for the regressor, then for the classifier) mainly confirm what we already knew. Greens and putts are important for scoring low, while chips and putts are generally the cause for my blow up holes.

Simulation

Now that we have a decision tree to (somewhat) accurately predict score, we can build a `Monte Carlo` simulation to see what our scoring distribution should look like, then adjust input parameters to see what we can do to actually lower our expected handicap.

Before we do this, we need to understand how handicap is calculated. Lots of golfers have trouble with this, so please enjoy this ChatGPT created explanation:

📉 What Percentile is My Handicap?

Handicap rule: A golf handicap is calculated as the **average of your best 8 scores from your last 20 rounds**.

We want to know:

> What **percentile** that score would fall into if your scores follow a normal distribution?

📐 Assumptions:

- Scores follow a **normal distribution** with mean (\mu) and standard deviation (\sigma).
- Your handicap is based on the **average of the lowest 8 scores** out of 20.

📊 Expected Value of Best 8 of 20 Scores:

The average of the **lowest 8 out of 20** values drawn from a standard normal distribution is approximately:

`MATH DOESN'T RENDER WELL IN MARKDOWN. DON'T WORRY ABOUT IT.`

📈 Percentile Calculation:

`AGAIN, DON'T WORRY ABOUT IT. THE ANSWER IS:` **Approximately 17.2%**

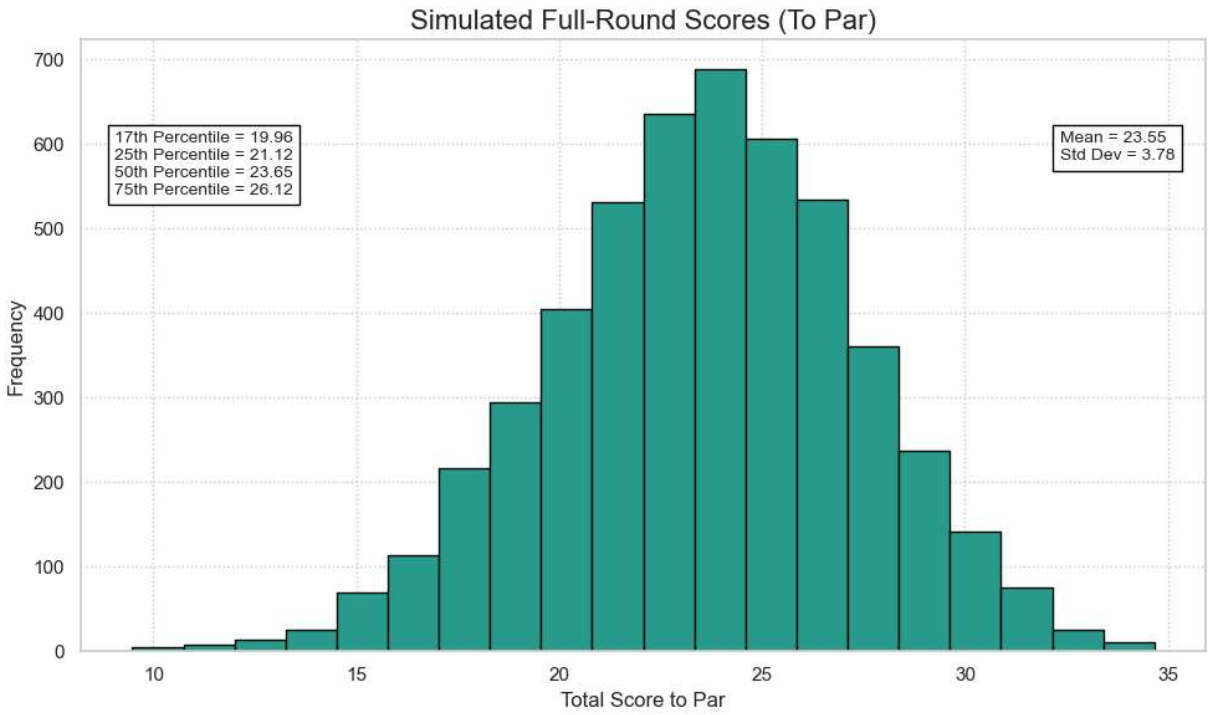
✅ Final Answer:

> Your handicap score represents roughly the **17th percentile** of your full scoring distribution.

This means:

- You play **to or better than your handicap** about **17% of the time**.
- It is a measure of **potential**, not average performance.

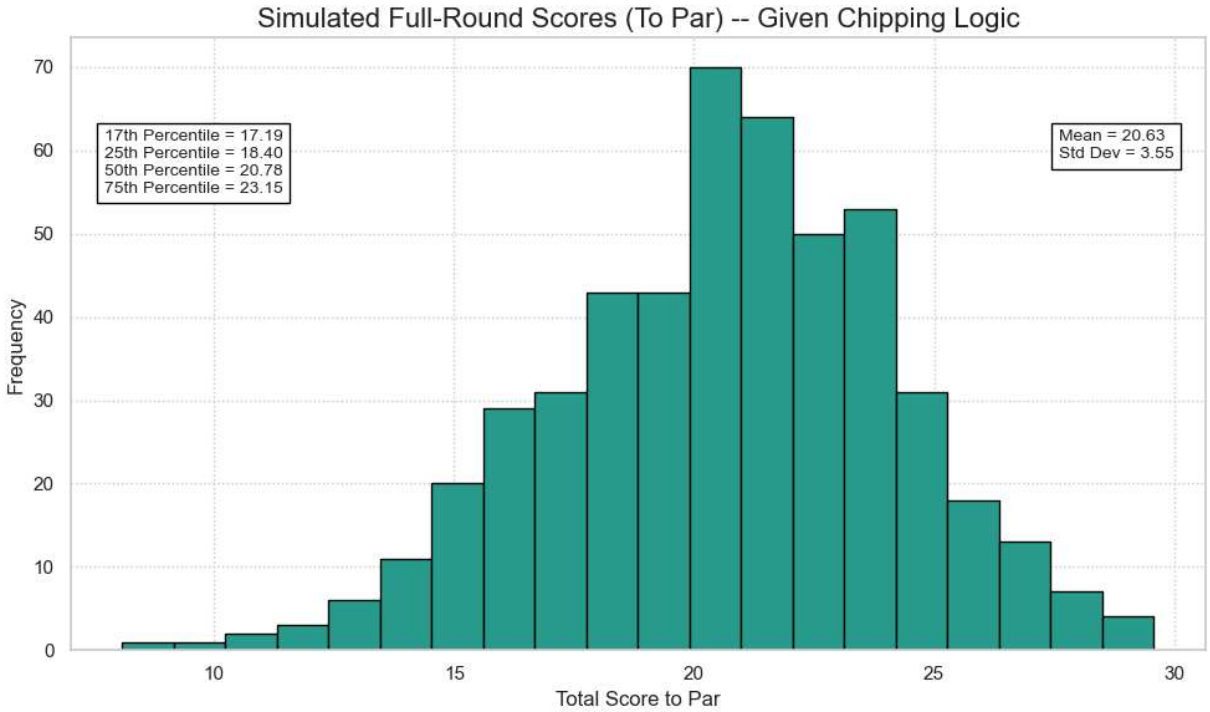
Ok. So in our simulated scoring distribution, we need our 17th percentile to be 15 strokes over par in order to reach my goal.



This is my simulated scoring distribution as it currently stands, assuming I play about average according to my last 20 rounds. My expected handicap is slightly lower than it really is, but in general this is a solid model that accurately represents reality.

Now, let's use some of the key influencers of my score to create **reasonable goals** to attempt to **lower my scoring distribution**.

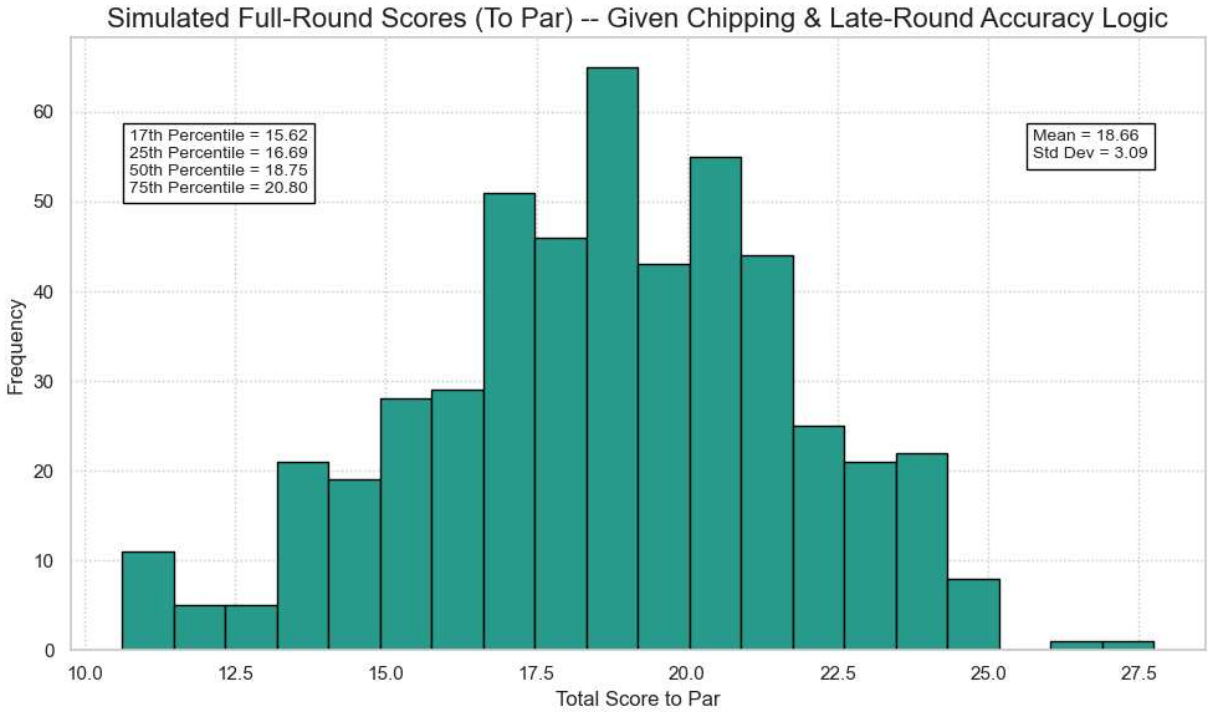
First, we will improve my chipping. Let's assume that when I miss the green, I only double-chip 10% of the time. When I hit a wayward tee shot with no chance to hit the green on my approach, I'll be allowed to double-chip 25% of the time.



Alright! This fictional world is getting better. If I focus on my chipping and reach those goals listed above, my expected average score will drop to approx 20 over par, and I will be about a 17 handicap.

Let's keep going. We know that I play worse at the end of the round. What if I set a goal to keep my tee shot in play at least 80% of the time on the last 6 holes? This means, limit those holes to 1 instance of having no chance to hit the green.

I am also adding in the very reasonable goal of no 4 putts or worse.



And there we have it. Three simple goals:

- limit double chips (with some grace on holes with bad tee shots)
- No more than 1 ball out of play for the last 6 holes of the round
- No 4 putts

If I do this, my expected average score drops to around 18-19 over par, and my expected handicap will be 15.

Conclusions

Thank you for joining me on this journey! I will try and keep this page updated as I continue to play, and hopefully as my scores drop (but you never know). Please feel free to reach out to me with any questions or comments. Go low, everyone!