Professional Golf Association Exploratory Data Analysis

Intro to Data Science Midterm Project:

Matthew Saxby & Mason Holland

Introduction

For 36 weeks out of every year, golfers from around the globe compete in the Professional Golf Association's tournament events in an attempt to win millions of dollars and gain an equal amount of fame. These golfers are willing to do whatever they can to gain an edge over their opponents in their golf game. We wondered, much like many professional golfers today, what are the most important aspects of the game of golf when you are trying to improve your average score and ranking? Can we help these professionals practice more efficiently and gain a mental edge on their competition?

When playing golf, there are usually 3 stages to each hole you play: driving, approach, and putting. Our analysis of the data will focus on these 3 stages and use the others less often. We have collected data from the 2022-2024 PGA tour seasons and for each year we have 8 different datasets which contain information regarding different statistics about each golfer that participated in each year. Our data is collected directly from the PGA tour website where they allow csv downloads for each of the datasets (*Golf Stat and Records | PGA TOUR*).

When approaching this task, we wanted to ask questions that we thought would be relevant to the PGA tour members, so we settled on asking 2 major questions: Which type of golf shot affects score and rank the most? Additionally, is it more important to be a powerful golfer or an accurate golfer?

When approaching the first question, we needed to compare all three types of shots: driving, approaching, and putting. We therefore decided we wanted to compare how players' overall rank and average score is reflected in how many shots better than the competition they had with respect to its shot category, otherwise known as "shots gained in [category]".

For the second question about accuracy vs power, we decided to compare the most accurate players and most powerful players. We found the most accurate players by analyzing the percentage of fairways each player hit, and the percentage of greens reached in regulation. To find the most powerful players, we looked for the players with the furthest average driving distances on the tour.

With the answers to these questions, we are hoping to be able to educate professional golfers about which elements will most effectively increase their rank in the PGA, and decrease their average score. We hope to be able to educate them based on the trends in the PGA over the last 3 years.

Data Description

Using the PGA website was helpful because we know that the data is trustworthy and consistent across the various years that we are working with. Some problems that we ran into were that we needed to merge the datasets that we were working with. Since we worked with 9 different data sets from each year and 3 different years, we worked with 27 different files.

We decided that we needed to merge all of our files from each year together, and we also created a super table with all the years merged into it so we could do an analysis between different years more easily. Because we were merging between different years, we also introduced NaN values into our dataset where some golfers did not compete in each differing year.

The following table shows the data structure for a year of data, with all 9 datasets from that year condensed into one table. It will cover each column with a description of the data and its data type.

Column Name	**Description**	**Data Type**
PLAYER	Player name	String
RANK	Player Rank	integer
AVG_Birdies	Average number of birdies per round	float
# OF BIRDIES	Season total number of birdies	int
AVG_Driv	Average driving distance	float
TOTAL DISTANCE	Total season driving distance	int
TOTAL DRIVES	Total number of drives	int
Avg Score	Average score each round	float
TOTAL STROKES	Season total strokes	int
TOTAL ADJUSTMENT	Season average score adjusted by course difficulty	float
AVG_SG:Tot	Average total shots gained	float
TOTAL SG:T	Season total shots gained	float
TOTAL SG:T2G	Total shots gained tee to green (driving)	float
TOTAL SG:P	Total shots gained putting	float
TOTAL SG: PUTTING	Total shots gained putting, based on a different amount of rounds as TOTAL SG:P	float
MEASURED ROUNDS	Total number of rounds played in a season	int
AVG_SGAppr	Average shots gained approaching green	float
TOTAL SG:APP	Season total shots gained approaching green	float
%_GIReg	Percentage of greens hit in regulation	float

Column Name	Description	Data Type
GREENS HIT	Season total greens hit in regulation	int
# HOLES	Number of holes played during the season	int
RELATIVE/PAR	Relative score to par	float
AVG_PUT	Average shots gained putting	float
%_Scram	Scramble percentage	float
Overall_Acc	Overall accuracy variable (Created)	float
AVG_SG_Driv	Average shots gained driving (Created)	float
PAR OR BETTER	Season total pars or better	int
MISSED GIR	Total greens hit in regulation	int

While we have collected a lot of data, it is pretty condensed. In each separate year, we are working with 170 rows and 27 different columns, each detailing a different variable impacting the game of golf.

We do have 3 different years, so as we merged our datasets, we began to introduce some anomalies where different golfers did not compete each logged year, and where the same golfers showed up multiple years in a row.

Many of these anomalies did not create major issues because of how we decided to combine all of the different years of data. We dropped all the golfers that did not show up in all 3 years and we also dropped columns that were not relevant for these golfers over the years. This allowed us to see variables like shots gained for different categories over 3 years. Which allowed us to see how professional golfers have changed their own playing style throughout the data collection time frame.

Data Cleaning and Preprocessing

The data used was sourced directly from pgatour.com. The PGA, or Professional Golf Association has provided statistic packed CSV files from the previous years. Unfortunately, they do not provide any compiled data, which meant that for each major statistic, we had to download a different CSV file. This in total meant that for each of the 3 years we wanted to conduct research over, we needed 9 CSV files.

We read in every file, and merged each of them into a combined CSV file for each year. We dropped unused columns before we merged them as several of the files that made up each year's data contained very similar information, with differing major statistics. The columns we dropped were "MOVEMENT","Player_ID","TOTAL_ROUNDS", and "RANK" for 8 of the 9 files per each year.

We filtered and combined the files for each of the years by using the merge function, and we merged on the player name, using suffixes to differentiate between similar columns to make our sets more readable. These suffixes ended in the last 2 digits of the year. We renamed all columns that were mixed up in the exchange, a problem we discovered when we were double checking to make sure our merge had gone smoothly. We decided to set the index of each year's dataframe to the player name so we had an easier time reading it.

This led us to the most challenging part of the merge, which was creating a merged data frame for all 3 years. This was moderately difficult due to the fact that each season there are different players in the PGA, so we had to deal with many NaN values.

We decided to merge on Player name, which meant we were going to be discarding the rows of players that do not occur in all 3 years. This meant fewer data points, but by doing this we were given access to a dataframe that would make our information more easily analyzed. It was also challenging, because all 3 years contained exactly the same columns, which meant we could not simply merge all of the columns, but we needed to create suffixes for every year, which led us to having 98 total columns, nearly all of them sharing a name with 2 others.

We created 2 new columns which contained statistics for Overall Accuracy and Average Shots Gained Driving which were used to be able to answer our questions further down the road. To create the Overall Accuracy column, inside of each year, we multiplied the Percent of Greens Hit in Regulation (par) and Percent Driving Accuracy. This gives us our own scale to be able to measure the common percentage of fairways and greens hit by a player, giving us an ideal number that represents their overall accuracy. For the Average Shots Gained Driving column we simply took the Total Shots Gained Tee To Green (driving) column and divided it by Measured Rounds. This will help us have a more accurate driving statistic to analyze.

Question 1: *Which Type of Golf Shot Affects Score and Rank the Most?*

Introduction

The overall goal in golf is to take as few strokes as possible to get the ball into the hole. A golf course consists of 18 holes, each varying in distance. There are three main distance categories, each with a different expected number of strokes to complete the hole, known as "par." Par ranges from 3 to 5 strokes, depending on the hole's distance. This is how golf is scored: in relation to the overall par of the course, which is usually a total of 71 or 72 strokes across 18 holes. One might hear a typical professional golf score of “-2 (two under)” meaning a golfer shot two strokes below the total expected stroke count of the course.

Every hole begins with a drive, ideally landing the ball near or in the hole, though this is rarely the case. As a result, players often take an approach shot—a shot aimed at landing the ball on the "green," the finely cut area of grass surrounding the hole. This can sometimes take more than one approach shot to get the ball onto the green. If the approach shot doesn't put the ball in the hole either then, the player will then putt, usually requiring more than one attempt to get the ball into the hole.

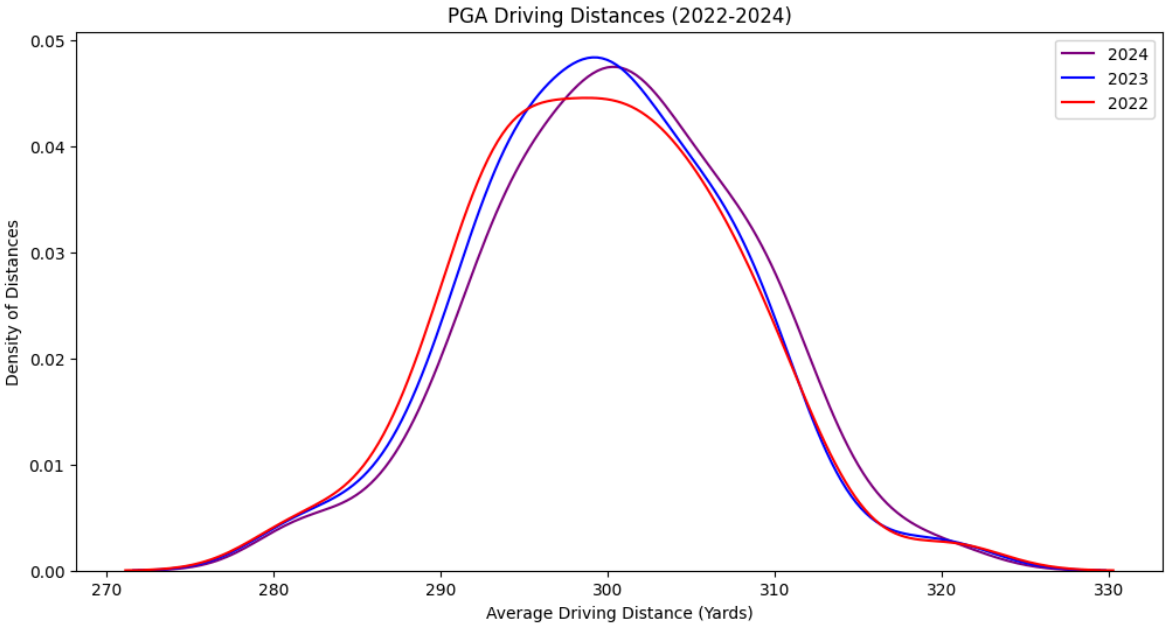
In this project, we are examining whether driving, approach shots, or putting has the most significant impact on a golfer's overall ranking and average score. These factors are often measured in terms of strokes gained in each respective category, meaning how many strokes a golfer performs better than the average competitor in that area of play. Average shots gained is how many shots they gain on their competition on average each round. Total shots gained, is the amount of shots they gained on their competition throughout the entire season.

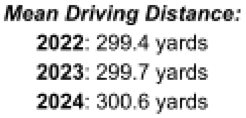
Method

Driving Analysis

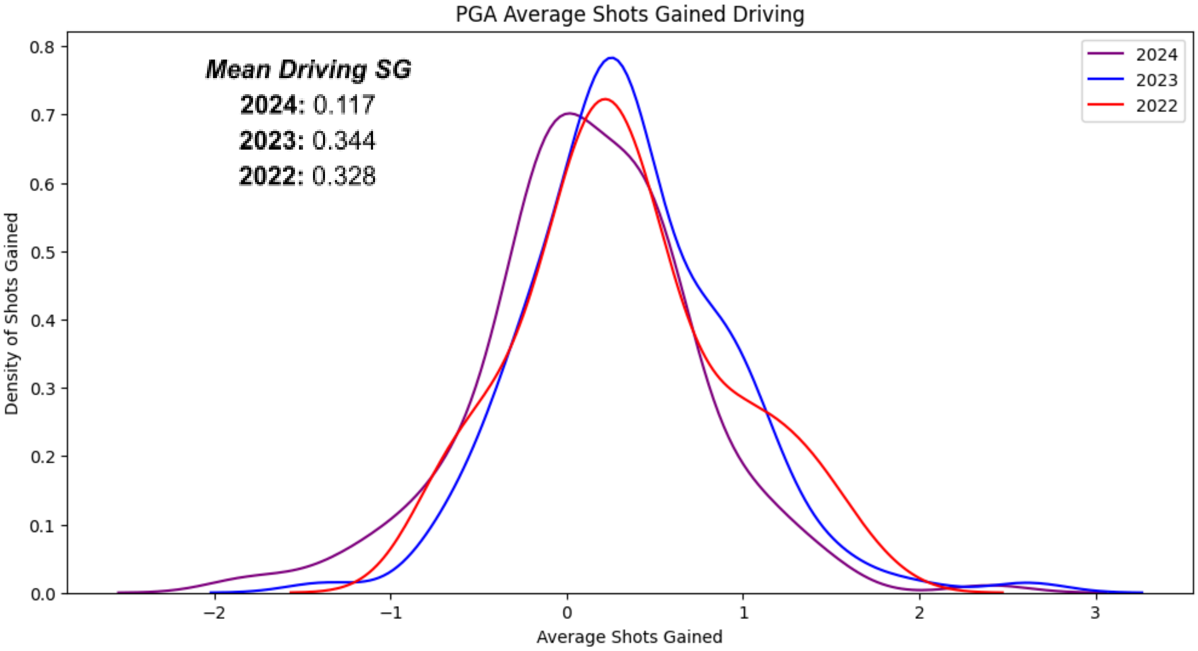
We began by analyzing the driving data from the years 2022, 2023, and 2024. We worked mainly with the variables (written legibly): Average Driving Distance (Yards), Average Shots Gained Driving, Total Shots Gained Driving, Average Score, and Average Rank. We wanted to look into how driving distance has changed over the last few years. This is a graph of the average driving

distance of each PGA golfer in the years 2022, 2023, and 2024.



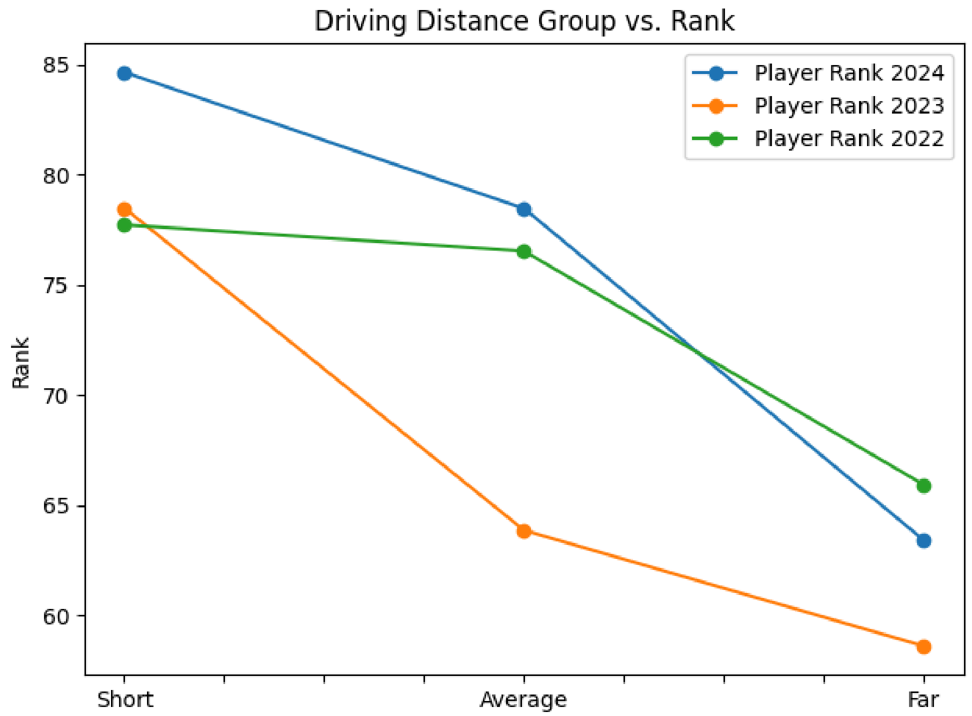

Mean Driving Distance:
2022: 299.4 yards
2023: 299.7 yards
2024: 300.6 yards

This data led us to see how driving distance has been relatively static over the 3 most recent years. With this information, we now wanted to know more about the average shots gained driving in addition to the distance. We chose to plot these in a very similar manner, so we can compare how the shots gained driving has changed over the last 3 years. Here is the graph to display this.

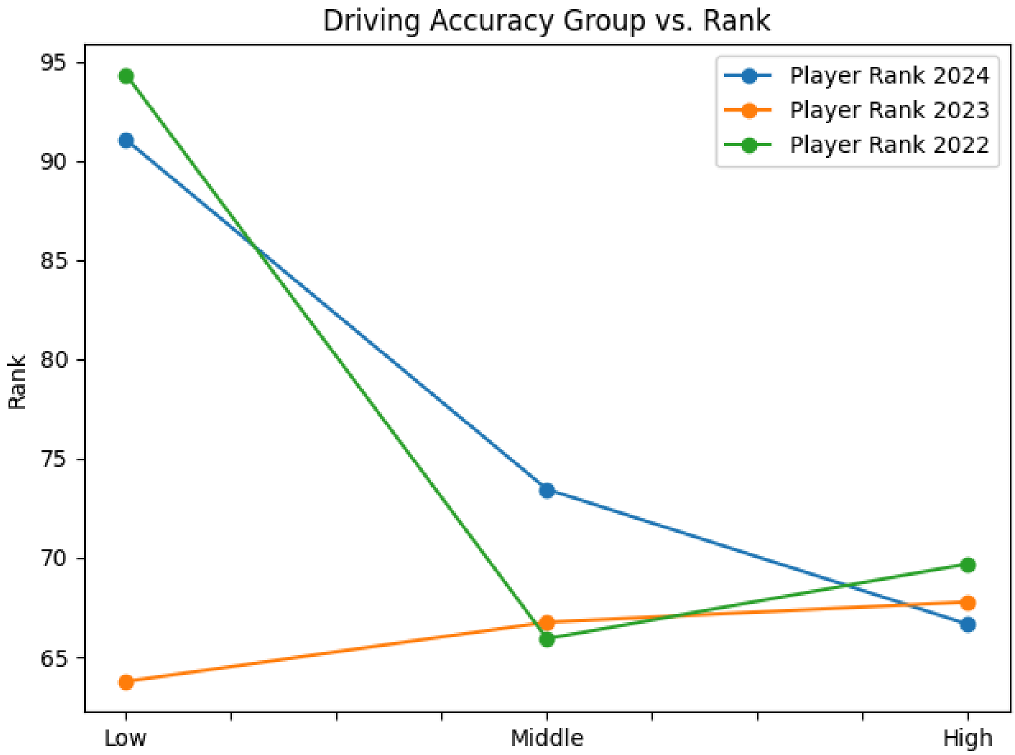


This data shows us that the average shots gained in driving has decreased slightly over the last three years, which we can confidently assume means that as average driving distance increases over time, so does the difficulty to gain a shot over another player in this category, as we can see that Average Shots Gained Driving, and Average Driving Distance (Yards) have an inverse relationship over the years.

We decided to break these distances into arbitrary groups to compare how distances compare over the years. Inside each year we grouped the average driving distance for a player into thirds, and assigned the first third to “short”, second third to “average”, and the last third to “far”. We compared this data to the average rank of the players in each of the driving distance groups, here is a chart that shows our results (the lower the rank the better). Doing this allows us to see how grouped distances relative to each year compares to other years, along with if the average rank for each group is consistent across years.



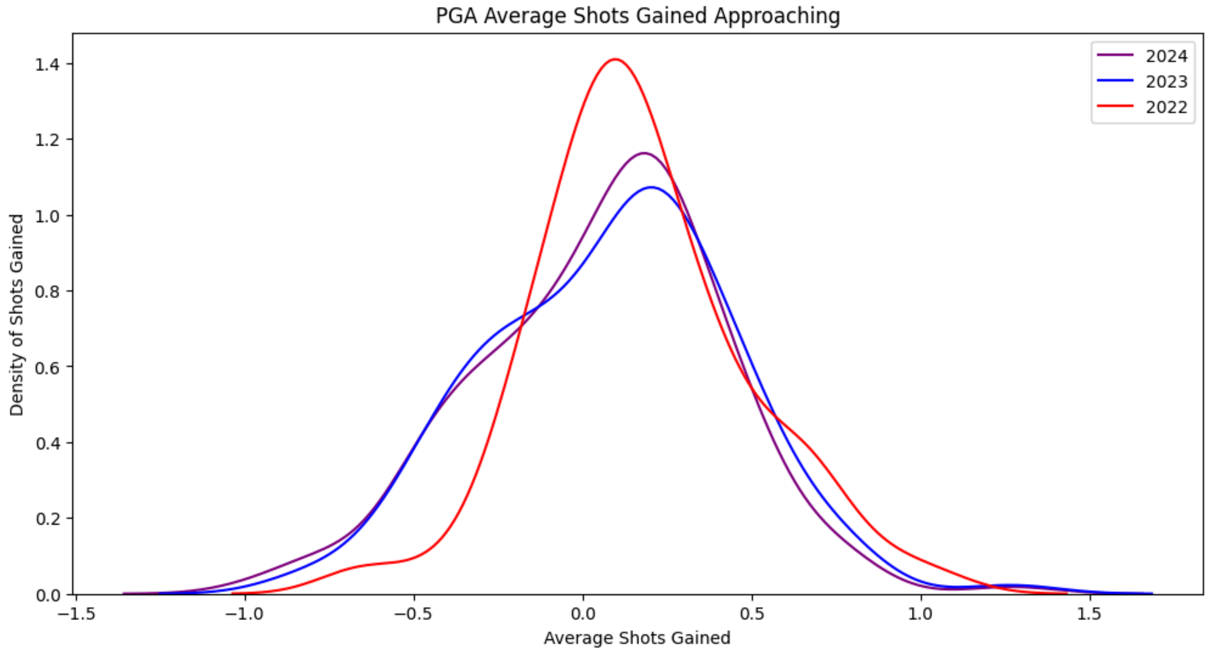
This graph shows that as average driving distance increases, so does the rank of the golfer. This is consistent across the 3 most recent years.We now wanted to use this exact same comparison but against average score in a round as opposed to the rank compared to other golfers. Here are our results.

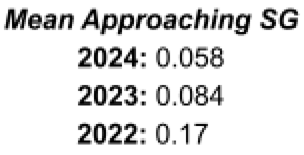


This seems to indicate that there is a clear relationship between driving distance and average score that shows that as a player's driving distance increases. Player scores tended to decrease, therefore improving their ranking.

Approach Analysis

We wanted to gain a proper view of the "shots gained approaching" distribution, so we plotted the trend line for each year next to one another to gain a better understanding of how shots gained approaching has changed over the last 3 years.




Mean Approaching SG

2024: 0.058

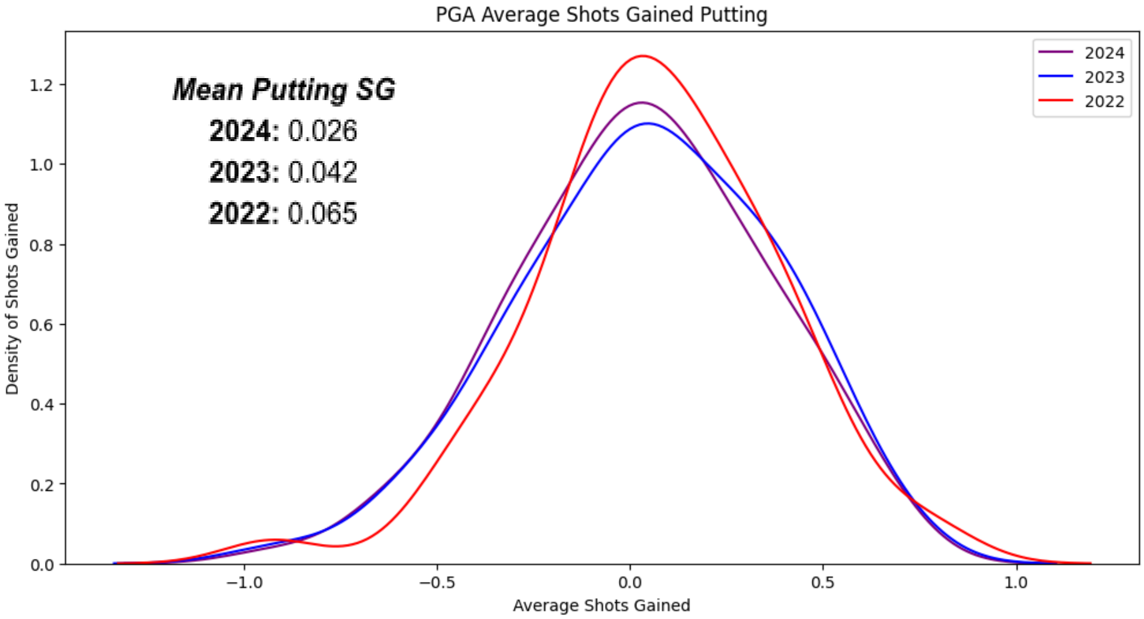
2023: 0.084

2022: 0.17

The graph above shows us that the mean is just a hair above zero, and nobody tends to have an average above 2 or below -2. We also see that the average shot gained approaching has also decreased over the last three years, our assumption is that the average golfer's approach shots have improved slightly, but analyzing that is not the aim of this report currently.

Putting Analysis

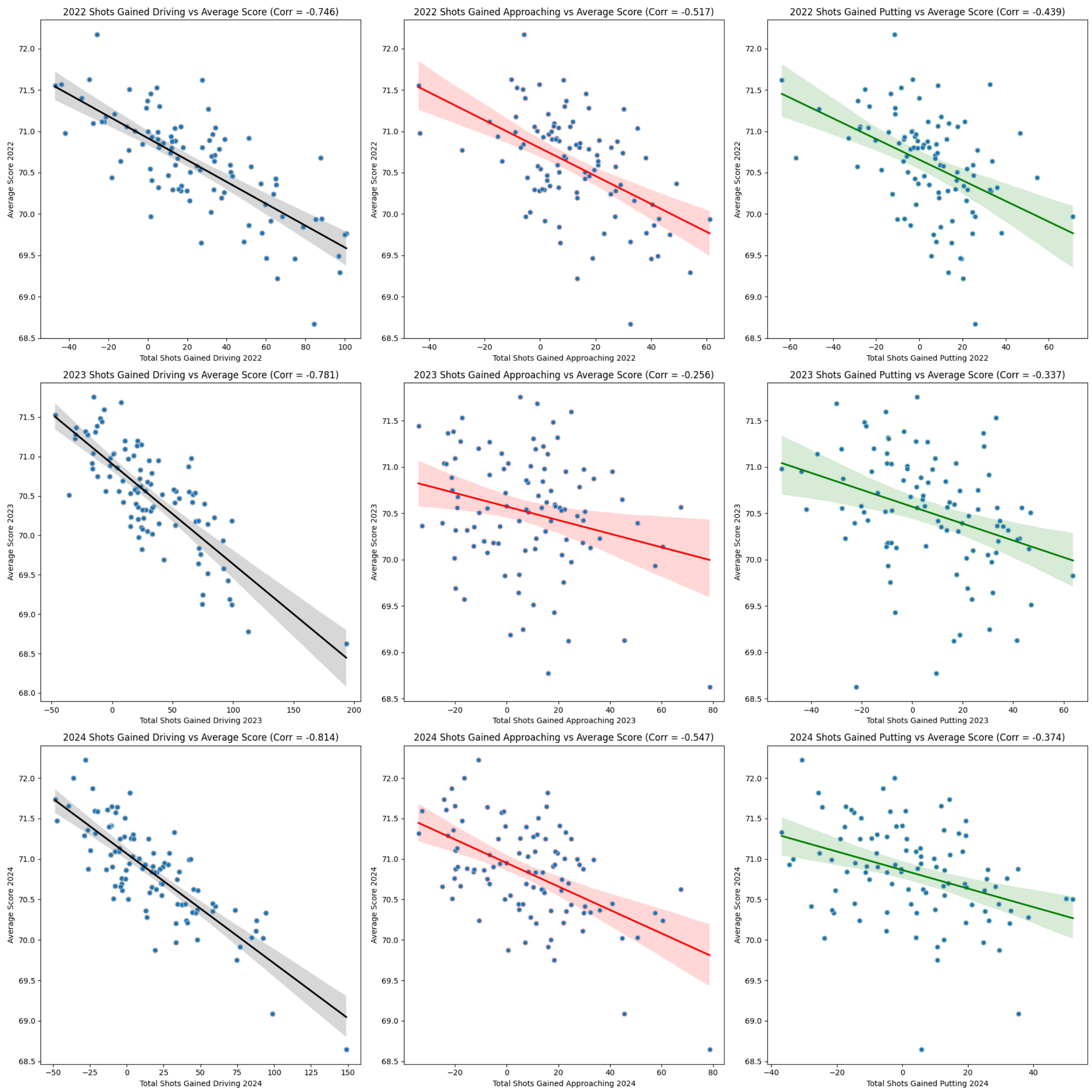
We wanted to gain a proper view of the "shots gained putting" distribution as well, so we plotted the trend line for each year next to one another to gain a better understanding of how shots gained putting has changed over the last 3 years. The graph is below.



This graph shows that the mean is just a hair above zero, very similar to approaching, and nobody tends to have an average above 1.5 or below -1.5. The average mean is decreasing, similar to the other average shots gained variables, but this decrease is incredibly slight, so it is difficult to be confident there is a trend with only 3 years of data present.

Driving vs Approaching vs Putting:

To begin, we analyzed (variables written legibly): Total Shots Gained Driving, Total Shots Gained Approaching, Total Shots Gained Putting, and Average Score. By plotting each variable against the year, we showed the relationship between all of the variables. We also calculated the average correlation for each relationship to show how driving, approaching, and putting shots impacted the average score. These correlations are negative because a lower score is better. In summary, Each point represents a single golfer. The Y-axis is their average score, and the X-axis is this golfer's total shots gained in the given category of shot.

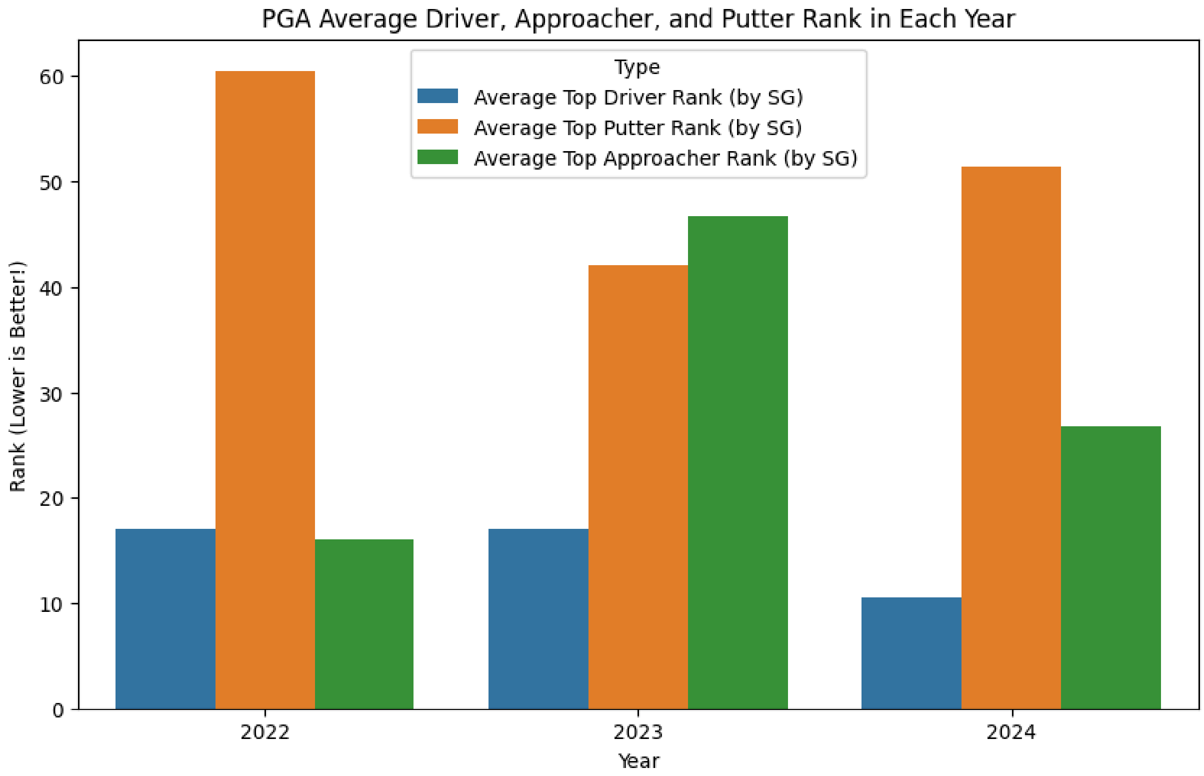


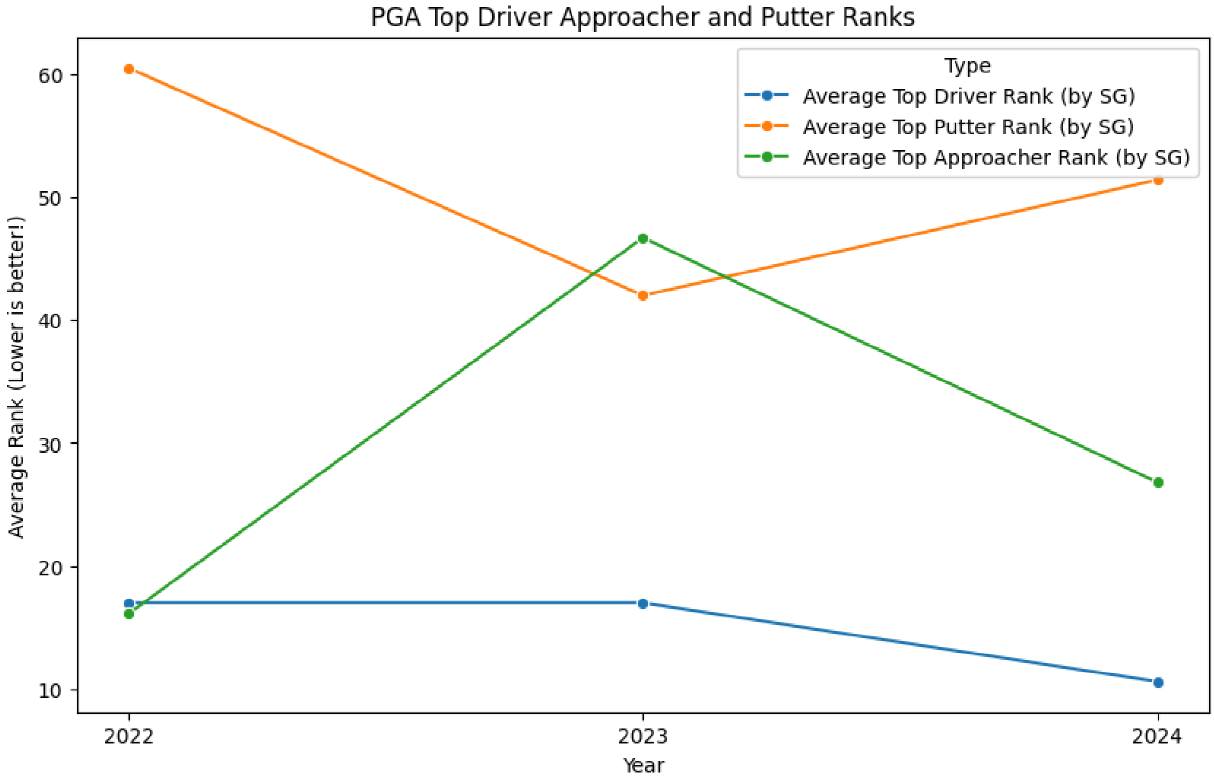
Below is a summarized table with the data presented above. We have gathered the correlation between the total shots gained of each type of shot and the average score for that player.

	Year	**Total Shots Gained Driving**	**Total Shots Gained Approaching**	**Total Shots Gained Putting**
Average Score	**2022**	-0.746	-0.517	-0.439
Average Score	**2023**	-0.781	-0.256	-0.337
Average Score	**2024**	-0.814	-0.547	-0.374

These correlations show us that gaining shots in any of these categories positively affect a golfer’s average score. But, it clearly shows us, more importantly, that over the course of the last 3 years, shots gained in driving have had the strongest correlation with a lower average score. Shots gained in approaching and shots gained putting have a positive effect towards decreasing a golfer’s score, but affect score roughly half as much as driving does.

We wanted to further analyze this data by comparing each type of shot, in regards to its average shots gained per player, and visualize how it affects Rank. We decided to take the 10 best golfers in each type of shot, based on total shots gained, and take the average of their PGA ranks. Both a bar chart, and a line plot are used below to display this. We chose the top 10 because we wanted to visualize only the best in each category.


PGA Average Driver, Approacher, and Putter Rank in Each Year
Type
Average Top Driver Rank (by SG)
Average Top Putter Rank (by SG)
Average Top Approacher Rank (by SG)
Rank (Lower is Better!)
0
10
20
30
40
50
60
2022
2023
2024
Year


PGA Top Driver Approacher and Putter Ranks
Type
Average Top Driver Rank (by SG)
Average Top Putter Rank (by SG)
Average Top Approacher Rank (by SG)
Average Rank (Lower is better!)
10
20
30
40
50
60
2022
2023
2024
Year

We can see from the charts that the average rank for the top drivers was consistently better than the top approachers and putters across 2022, 2023, and 2024. This follows what we saw in the correlations between different shots and average score above.

Discussion

This points to the fact that driving is where we see professional golfers gaining the most shots against the competition. It is key to note that all of these areas are essential to becoming a golf champion, but on average, a player who has more shots gained in their driving will typically have a lower average score in any given professional golf round.

Question 2: Is it better to be a powerful golfer or an accurate golfer?

Introduction

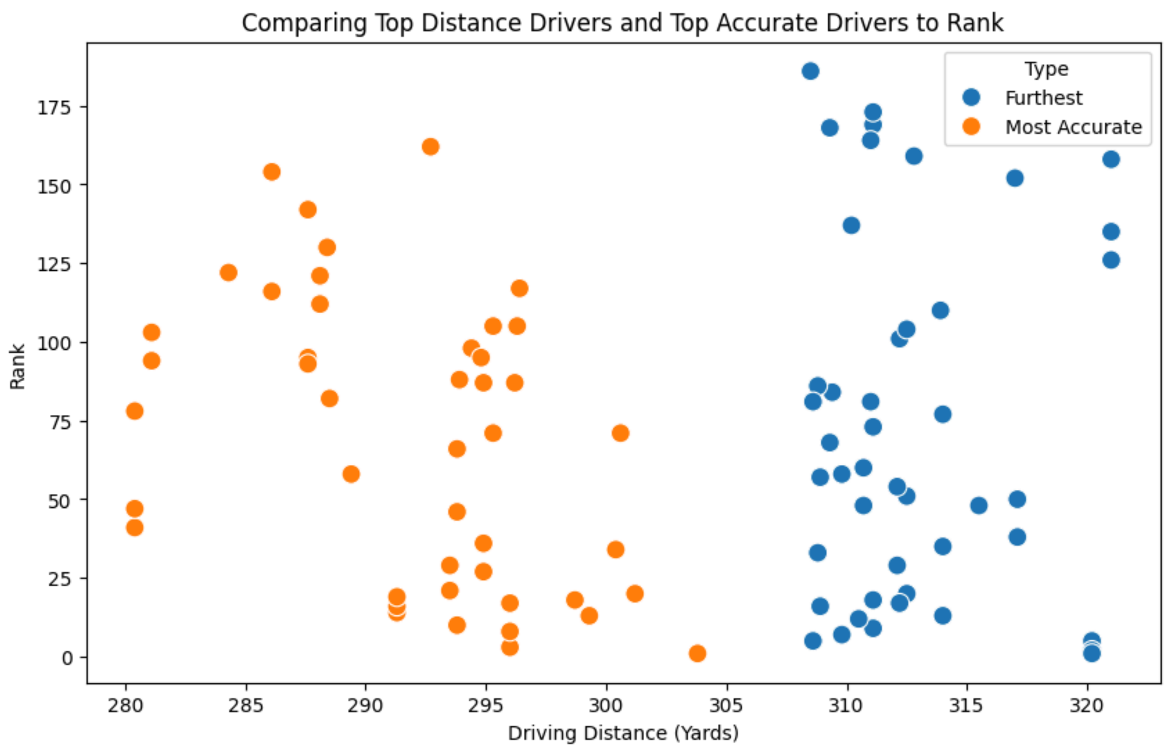
In golf, the ideal player would be able to drive the ball 330 yards straight down the center of the fairway with every swing, but this is nearly impossible. Many professional golfers face a trade-off between hitting the ball with maximum power, sacrificing accuracy, or focusing on control and sacrificing distance. We wanted to analyze this data to determine whether average score increases or decreases based on whether a golfer focuses on power or prioritizes accuracy.

First, to properly assess a golfer's strength or accuracy, we must select variables that reflect these qualities well. For strength, using driving distance is a simple and straightforward measure of power. Accuracy, however, is more complex to quantify. For this analysis, we used greens in regulation and fairways hit to measure accuracy.

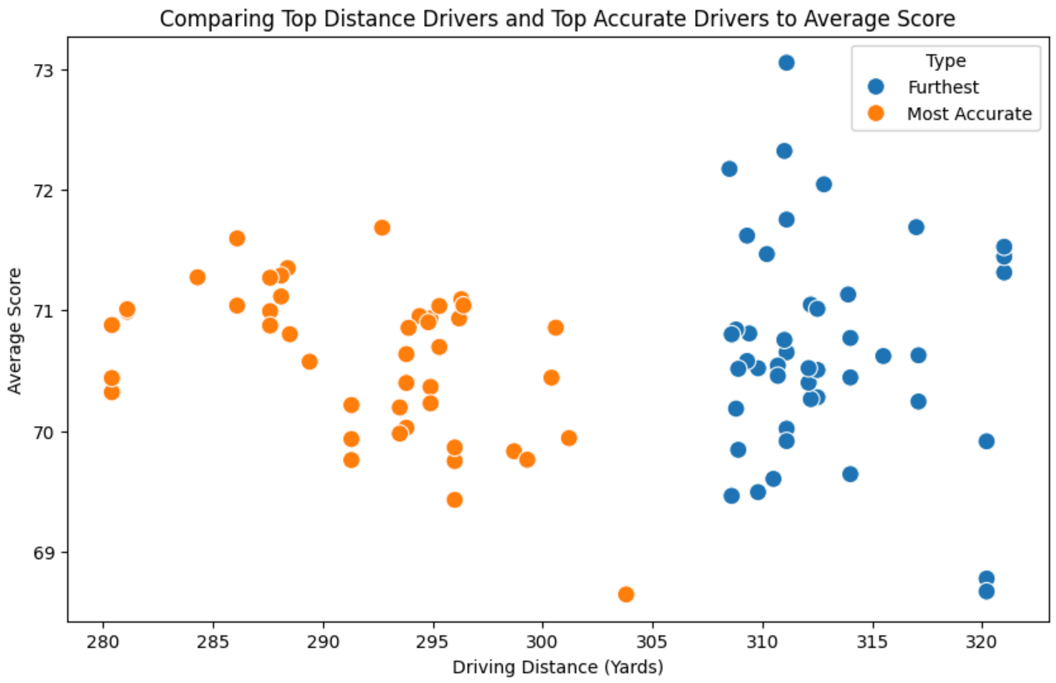
Greens in regulation refers to whether a golfer lands the ball on the green within the expected number of strokes. For a par 5, landing on the green by the third stroke or sooner counts as a green in regulation. For par 4 and par 3 holes, it is 2 and 1 strokes, respectively. Fairways hit measures whether a golfer hits the fairway with their tee shot otherwise known as their drive. A high percentage of fairways hit indicates accuracy, while greens in regulation reflects accuracy as well but also incorporates some influence from driving distance.

Method

The variables we used for this comparison were (written legibly): Average Driving Distance, Average Score, Rank, Driving Accuracy, and Overall Accuracy. The first step in creating visualizations was to separate the best drivers in terms of distance in addition to the most accurate drivers. We created these visualizations using the top 15 golfers in driving distance and the top 15 golfers in fairways hit (pure driving accuracy). We chose the top 15 golfers in each category in each of the last 3 years because we wanted to compare the most elite of both categories. We know the top 15 furthest driving golfers will be on the right side of this graph, but what we are looking for is if any of the most accurate drivers will also be some of the furthest drivers.



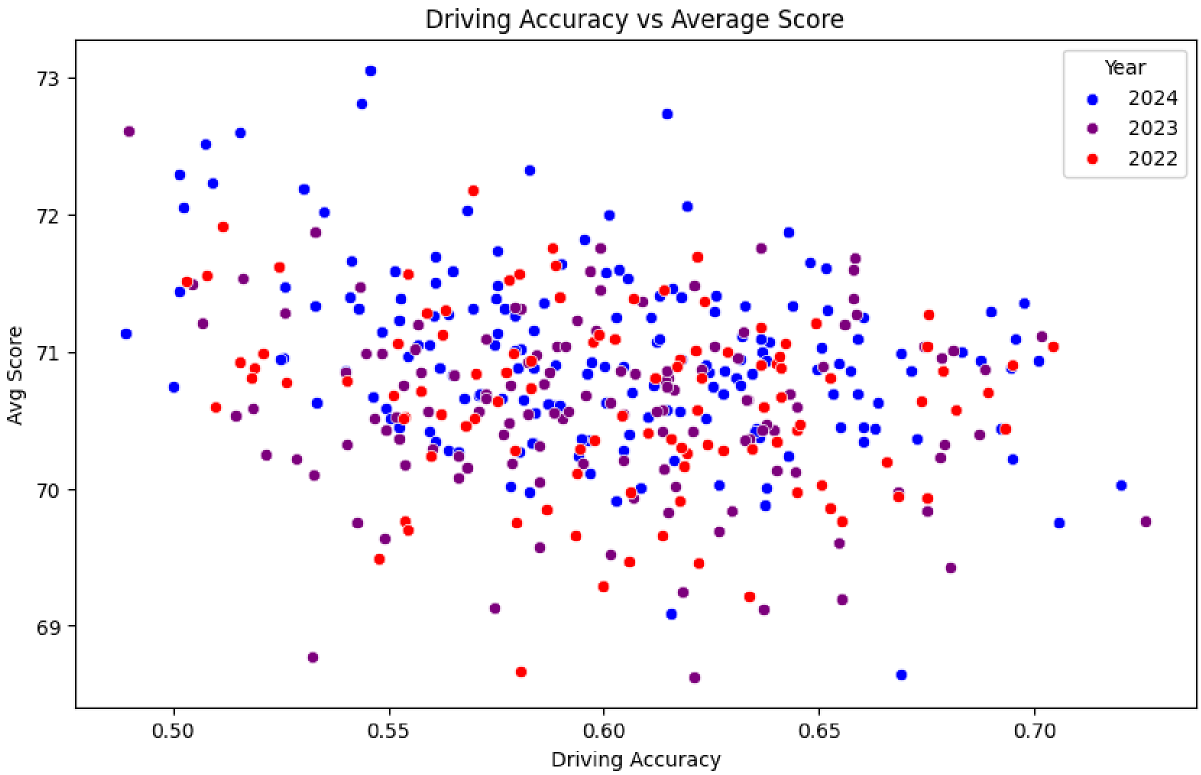
It is clear that none of the top 15 most accurately driving golfers over the last 3 years have ever also been in the top 15 longest driving golfers.



We decided to also investigate this same relationship but with Average Score instead of Rank, and there is no overlap between the most accurate drivers and the furthest distance drivers. All accurate drivers happen to be average or short length drivers.

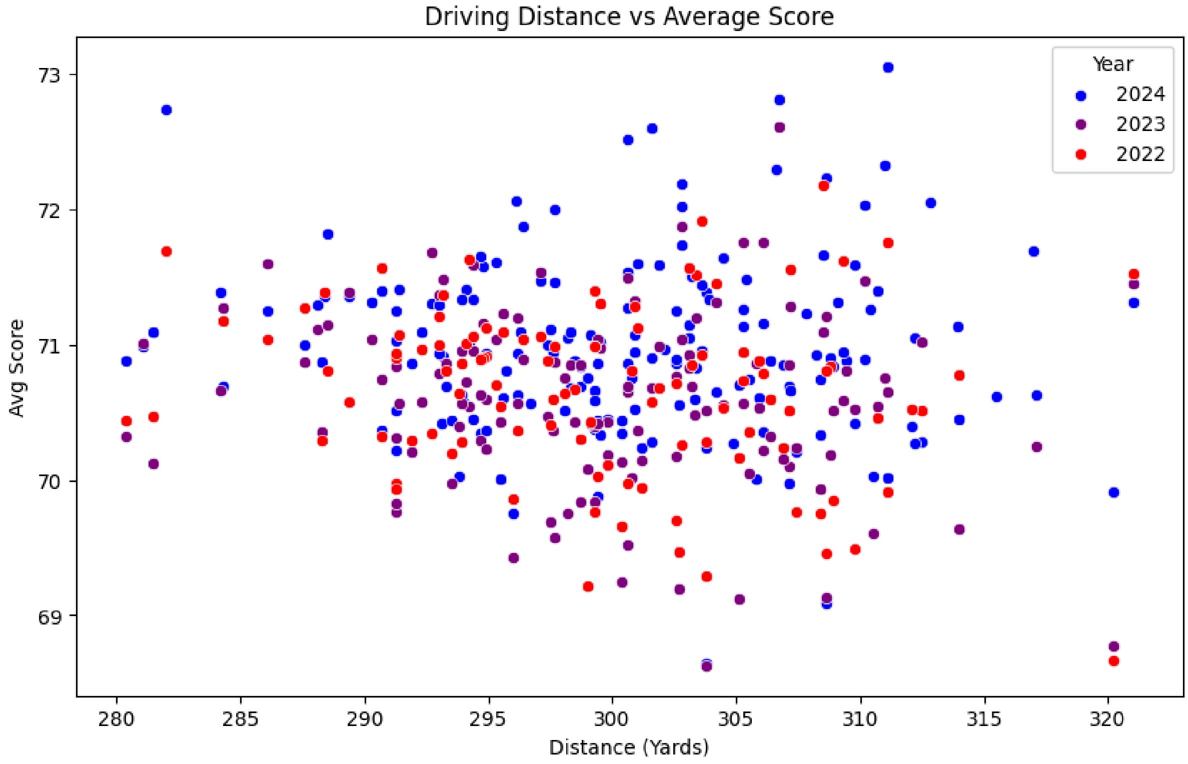
We found this incredibly interesting that the golfers that drive the ball most accurately in the last 3 years have never overlapped with the golfers who drive the ball the furthest.

We decided that we now wanted to investigate Average Score versus Driving Accuracy to see the relationship between the two, so we can compare it against Driving Distance.



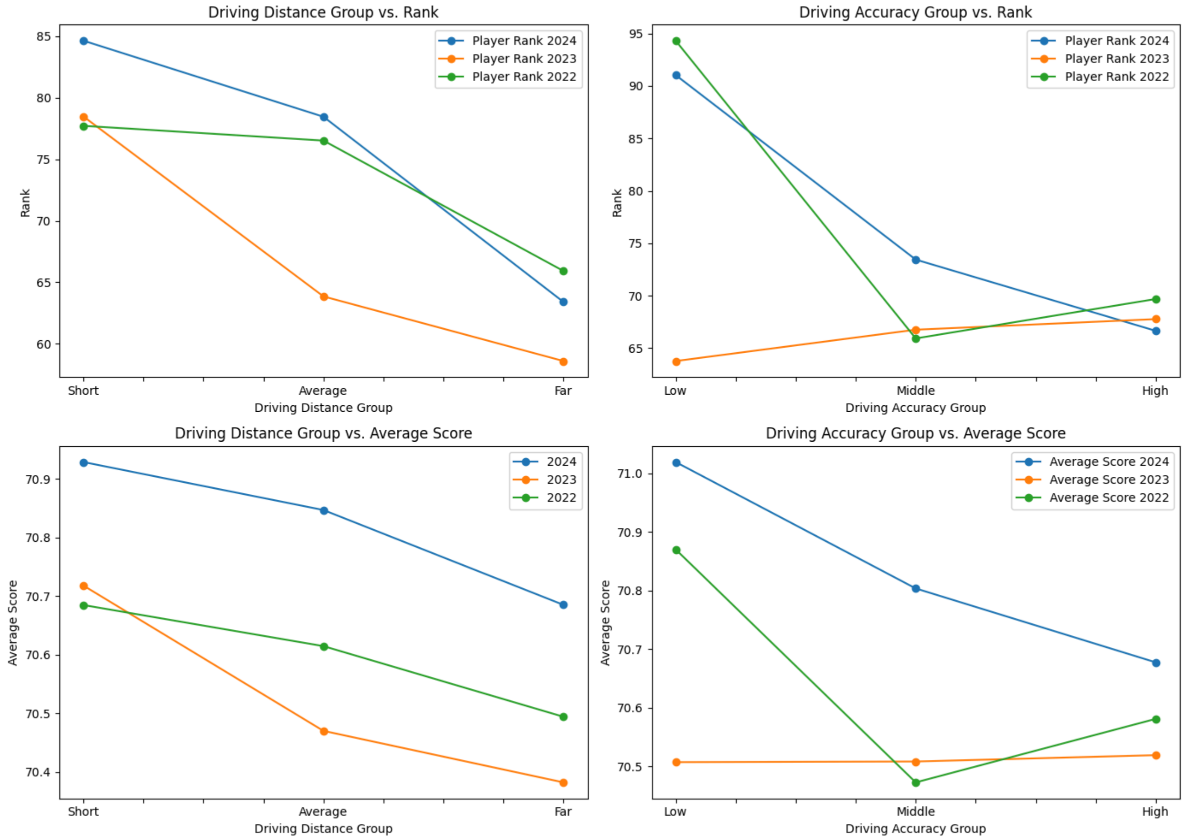
As driving accuracy increases, average score tends to decrease slightly, but the most accurate players seem to have slightly better average scores than the less accurate players. These variables have a correlation of -0.226, which indicates a relationship, but not an incredibly strong one.

This graph led us to question if this correlation between accuracy and score is greater than that of driving and score, so we created the same graph, but with driving distance instead of driving accuracy.



The correlation between average driving distance and average score in between 2022-2024 is -0.12 which is weaker than the correlation between overall accuracy and Average Score. This leads us to the conclusion that there is a slightly stronger relationship between Driving Accuracy and Average Score then between Driving Distance and Average Score.

We decided to break the Driving Accuracy variable into categories of low, middle, and high accuracy, and to compare this against both Average Score and Rank. This is so we can compare this against the driving distance’s cut into groups that we presented earlier.



These graphs show that while driving accuracy may have a stronger correlation towards decreasing average score and rank, this is inconsistent across all seasons. The data presented from the driving distance shows a much more steadily lowered score as drive distance increases.

Discussion

These results overall seem to sum up a partial but somewhat reliable result. Due to how important driving distance is to increasing overall shots gained, and summing up all of our data, we conclude that while very similar, the further drivers seem to provide a more consistently lower average score than accurate drivers do. It seems to be slightly more beneficial to be a more powerful player as opposed to a purely accurate player, as it more consistently shows to improve shots gained in addition to lowering average score.

PGA Tour EDA Conclusion

After wrangling, transforming, and analyzing data from the PGA Tour in the 2022-2024 seasons, we learned that there is indeed a key element to maximum success in professional golf and that key is driving distance. When we looked at which shot is the most crucial to predicting score and improving rank, we found that players gain an edge over their competitors most effectively by driving the ball further. We also explored whether being a more accurate player could be more beneficial than being a powerful one, but the results showed that consistency does not always improve results.

Recommendations

Given our findings, our recommendation for professional golfers would be to prioritize increasing their driving distance if they want to improve their ranking and lower their average score. We've found a direct relationship between longer drives and more shots gained from driving. This is significant because our data determined that gaining shots in driving has the most direct relationship to improving score and rank.

However, it's important to note that aiming for greater distance can introduce more variability. Typically, longer drives come at the cost of accuracy. Therefore, if a golfer is worried about maintaining consistency and avoiding falling out of the tour, we would recommend focusing on accuracy and hitting the fairways as that might be more beneficial. This approach could help them make more tournament cuts and boost their earnings, especially if they're struggling to consistently place in the top half of the field.

Discussion and Limitations

Having data across the three most recent years has given us brilliant insight into the professional golf world. We can see very clearly the importance of specific areas to the golf game. Because we were able to see direct correlations between variables, we were able to visually depict trends in the data which led us to our conclusions.

Our dataset contains lots of great data, but it is limited in the format we received it. Because this data came pre-formatted to show averages, shots gained, and other information, we were not able to create many new creative variables based on raw data from every shot in every tournament from every player. This ideal data would have taken significantly more time to process, but it would have given us the ability to analyze putting and approach distances, weather conditions, and other variables that play into a round of golf.

Works Cited

PGA Tour Website (Data Source): *Golf Stat and Records | PGA TOUR*

Matplotlib Documentation: *Using Matplotlib — Matplotlib 3.9.2 documentation*

Pandas Documentation: *User Guide — pandas 2.2.3 documentation*

Seaborn Documentation: *User guide and tutorial — seaborn 0.13.2 documentation*

Jupyter Labs Workspace: *Jupyter Notebook*

Contributions

Matthew Saxby	**Mason Holland**
- Conceptualization - Data curation - Formal analysis - Investigation - Methodology - Project administration - Software - Resources - Supervision - Validation - Visualization - Writing – original draft - Writing – review & editing	- Conceptualization - Formal analysis - Investigation - Methodology - Project administration - Software - Resources - Supervision - Validation - Visualization - Writing – original draft - Writing – review & editing

Acknowledgements

Thank you to Ellie Stamper for peer reviewing this report.