Analysis of PGA Tour Results and Player Statistics

Fitting a Linear Regression Model to Tournament Results

Kyle Arbide

SEAS 6401

Dr. Benjamin Harvey

December 7th, 2021

Table of contents

Abstract

As sports data becomes widely available online, the popularity of sports analysis also grows. Sports statisticians are constantly trying to improve their prediction models in an attempt to "beat the odds". This report is an attempt to become familiar with sports data and analysis and its application using PGA tour data and player statistics. It also uses these statistics to create a linear regression model that will predict tournament results. There were many steps in attempting to increase the validity of this model, resulting in a not so accurate regression model with an R-squared score of ~0.45. Finally, the result of this regression will be compared to other researched models and conclusions will be drawn on where to improve moving forward.

Introduction

Data analytics has revolutionized how players and coaches make decisions on the PGA tour. On top of the decisions players make on the course, the Tour uses ShotLink data to enhance the fan experience when watching the live broadcast. In this report, I hope to replicate some of these analyses by creating both interesting visualizations of exploratory statistics, and a more advanced model for performance prediction. The exploratory analysis consists of visuals that display the performance of players throughout the season, player comparison through distance matrices and spider plots, as well as a look at how individual courses played. The performance prediction will be conducted through a linear regression model that uses the player statistics from the exploratory analysis, as well as a few additional statistics calculated from the results data, to attempt to predict results at a given tournament. The dataset used for this analysis contains the results for all the tournaments from 2020, as well as the first few tournaments from the 2021 season.

Data Wrangling

Data Identification

Before being able to perform any type of analysis, I would need to obtain two different types of datasets: individual player statistics and individual player performances. Locating the former was simple. Kaggle has two datasets with recent statistics for players on the PGA tour. After combining these and keeping the players with a full list of stats, I was left with a dataset of 171 players and 29 different statistics that fall into 1 of 5 stat categories. These categories are:

Greens In Regulation: Describes how frequently the player makes ir to the green at least 2 strokes away from par based on a number of situations. Evaluates a player's skill in the fairways/middle game.

Tee Box: Describes different elements of a players driving/tee shots. Evaluates a player's skill off the tee/long game.

Putting: Describes a player's performance on the green. Evaluates a player's putting skill/short game.

Performance Based: Describes a player's performance in terms of previous results and scores. Evaluates a player's consistency and past performances.

Strokes Gained: Combined, numeric evaluation of the quality of a player's individual shots across all aspects of play. Evaluates a players overall skill level compared to the rest of the field.

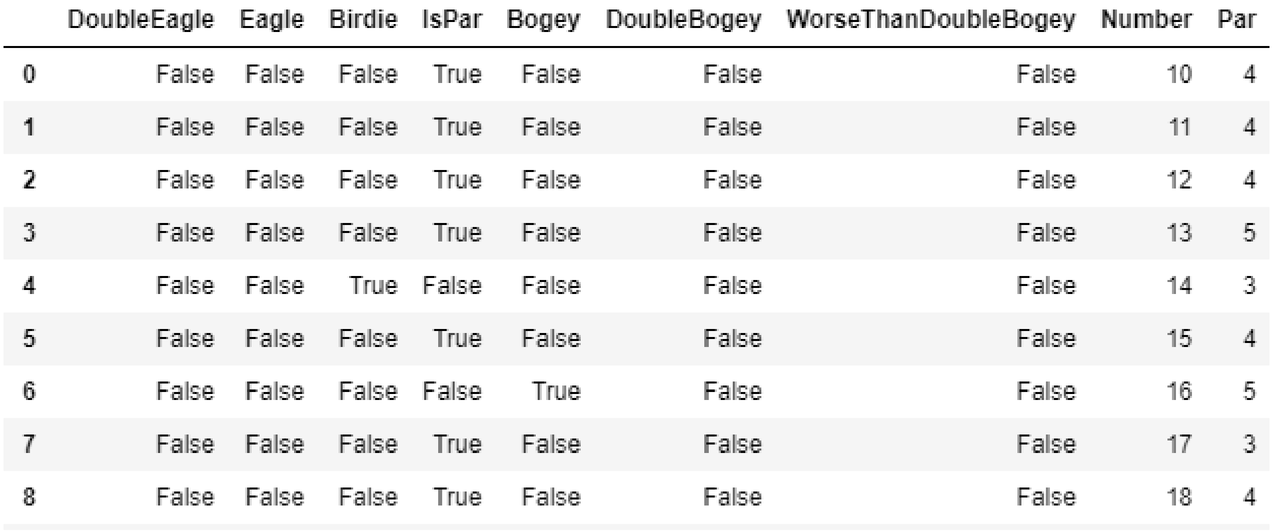
There was, however, some difficulty in accessing player results, especially for a large collection of tournaments/ the entire season. As companies start to recognize the values of granular tournament results for sports betting, many of the best data sets have been blocked by pay walls and subscription requirements. This would lead me in the direction of SportsData.io, a sports data API that offers a free trial of their result data for the last and current season.

Data Extraction

Sportsdata.io provides a free trial key to access their API and use a limited version of their comprehensive data. This allowed me to pull the player and tournament primary keys for each of recent tournaments and players from my statistics dataset. The situation gets more complicated when extracting the actual result data, as each API pull will only allow you to get a single tournament result for one player. For this, I created a loop that went through the lists of players and tournaments I wanted to extract and combined the results into one comprehensive dataset. The function also needed to identify a few key edge cases, including players not participating in one of the tournaments, or a player not completing any given round (potentially due to injury). The extraction is able to identify these issues automatically, as well as perform the aggregation required to have a comprehensive dataset of results.

Data Aggregation

Another major limitation in acquiring tournament results by player was the format in which SportsData.io, which can be seen below in *Table 1*.



	DoubleEagle	Eagle	Birdie	IsPar	Bogey	DoubleBogey	WorseThanDoubleBogey	Number	Par
0	False	False	False	True	False	False	False	10	4
1	False	False	False	True	False	False	False	11	4
2	False	False	False	True	False	False	False	12	4
3	False	False	False	True	False	False	False	13	5
4	False	False	True	False	False	False	False	14	3
5	False	False	False	True	False	False	False	15	4
6	False	False	False	False	True	False	False	16	5
7	False	False	False	True	False	False	False	17	3
8	False	False	False	True	False	False	False	18	4

Table 1: Raw Tournament Result Data

Results were given on a hole by hole basis, and with a Boolean data type that states "True" for the result from that hole. This meant I had to add 3 steps to the data extraction process. The first step isolates the result from that hole by pivoting the result columns. The second step converts the text result into a numeric (i.e., Birdie = -1) and the third combines the results across a round to output round by round scoring data for a tournament, which later can easily be aggregated into an overall tournament score.

This meant there were 3 datasets taken into the exploratory stage: the 29 player statistics mentioned earlier, a round by round summary of tournament scores by player, and the hole by hole results at a tournament by player.

Analysis

Exploratory Analysis

During the exploratory analysis, we are able to use visualizations to draw some initial conclusion about the tournaments and players in our dataset, as well as some general conclusions about performance across tournaments. Theses first visuals created used average round score to rank tournaments on difficulty and players on performance.

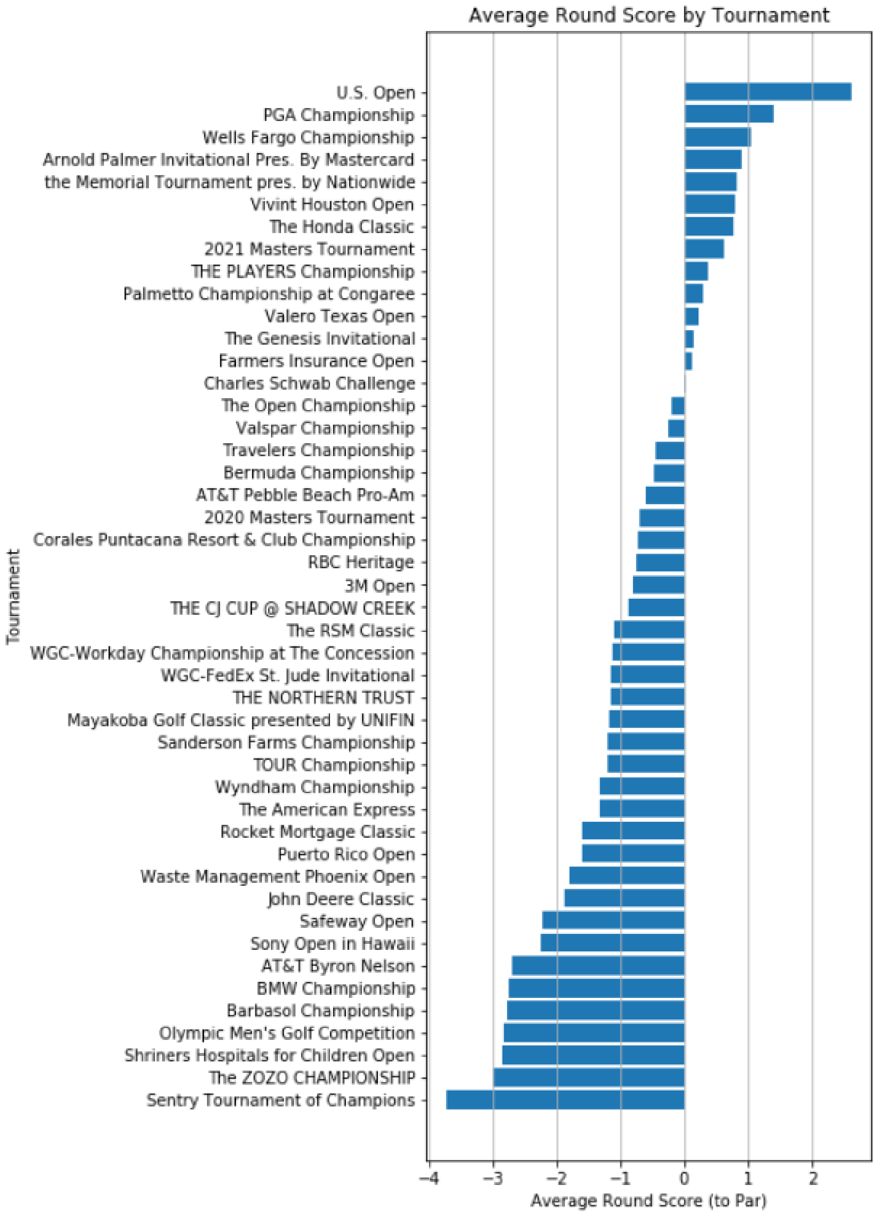


Figure 1: Average Round Score by Tournament

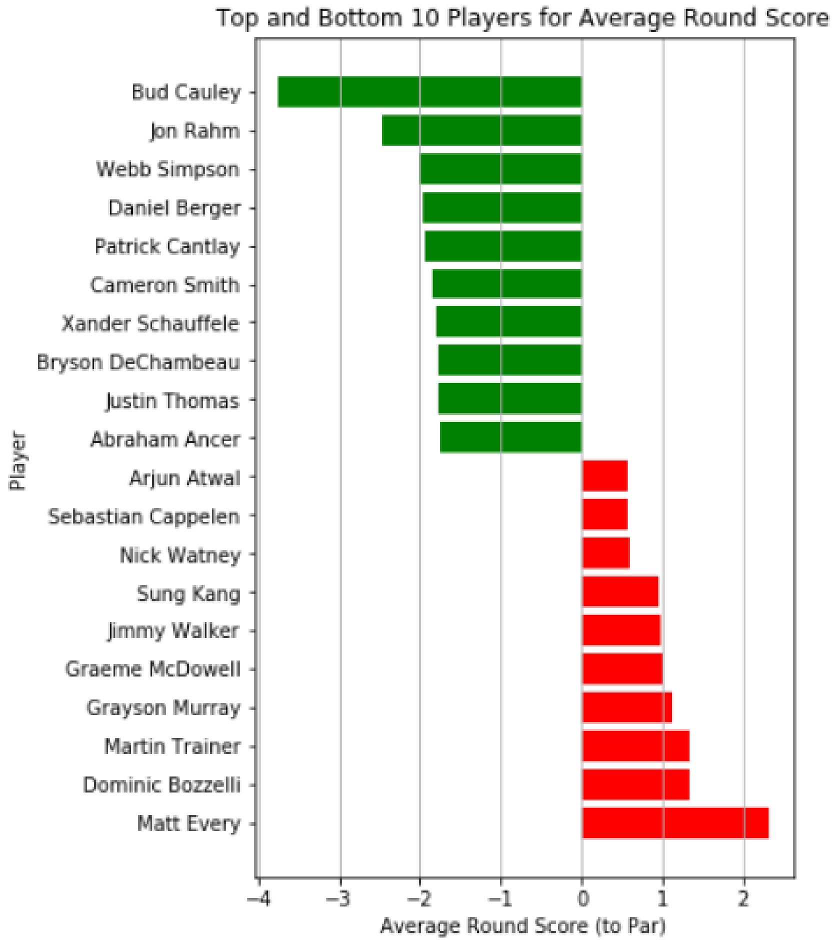


Figure 2: Top and Bottom Players by Average Round Score

Figure 1 and *Figure 2* allow us to draw some initial conclusions about the difficulty of tournaments and quality of players in the dataset. In *Figure 2*, we would suspect the players with the green bars were some of the higher performers for the season, and thus some of the best golfers. Similarly, we might conclude from *Figure 1* that the Sentry Tournament of Champions has the easiest course, and the US Open has the hardest.

The hole result dataset also opens the door to create some interesting analyses about hole by hole performance. The following visuals displayed in *Figure 3* and *Figure 4* provide some interesting insights about hole difficulty.

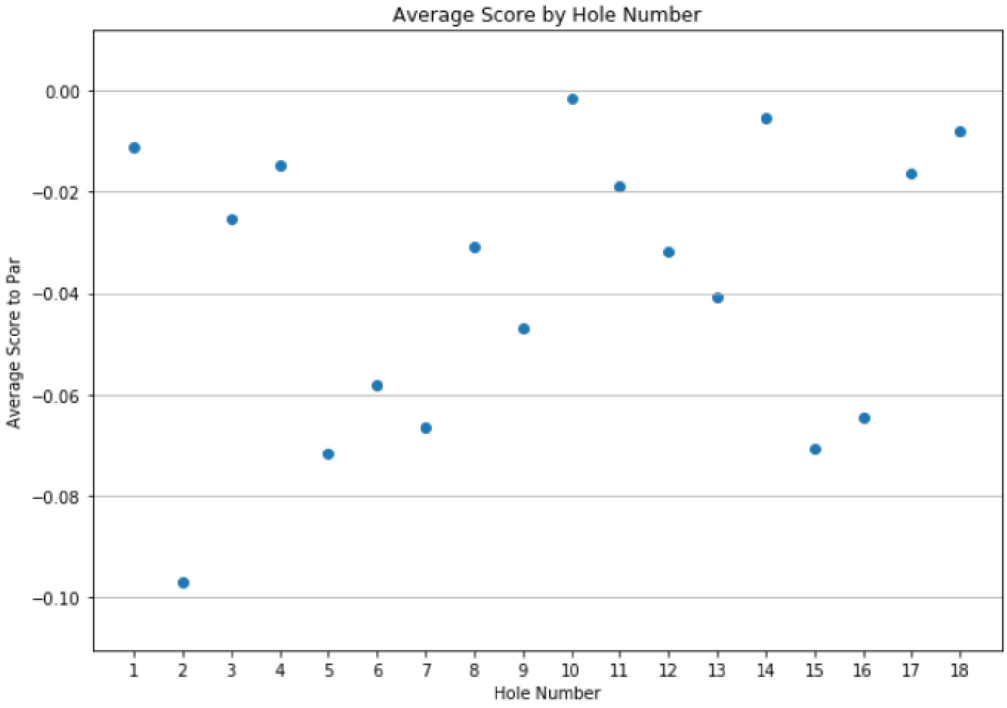


Figure 3: Average Score (to Par) by Hole Number

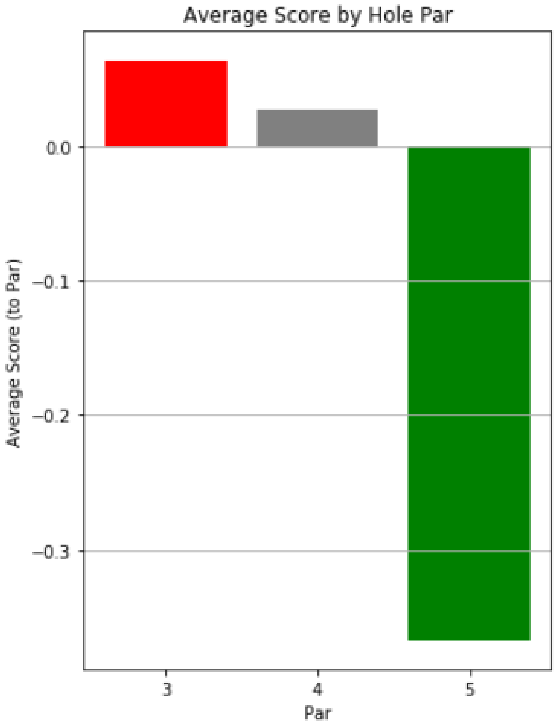


Figure 4: Average Score (to Par) by Hole Par

Figure 3 is the scatter plot of the mean score by hole number across all tournaments. The visual seems to suggest that hole 2's are generally the easiest/ lowest scoring while holes 10,14, and 18 are higher scoring. This same visual can be applied to an individual tournament, to see which hole on the course players generally struggled or found success with. *Figure 4* uses the same average hole score metric, but this time comparing the pars of the holes. Par 3s and 4s average a score above zero, while par 5s average a drastically lower score. This suggests players could look at par 5s as opportunities to make up score.

Comparison Metrics

Its common when watching broadcasts for sporting events to see player vs player or team vs team comparisons based on key stats. This can be very difficult to do as some player might excel in some categories and perform poorly in others. To perform some player comparison of my own, I created distance matrices based on tournament for both players and tournaments. This should group together players that performed similarly across all tournaments, and tournaments that performed similarly across all players. Distance matrices were made using both Euclidian and cosine distances, with and without mean substitution for NA values (created when players don't compete in one of the tournaments). The result is a dendrogram which groups players and tournaments based on similar scoring results. Most of these relationships are pretty weak, but if we highlight the strongest relationships some groups start to form.

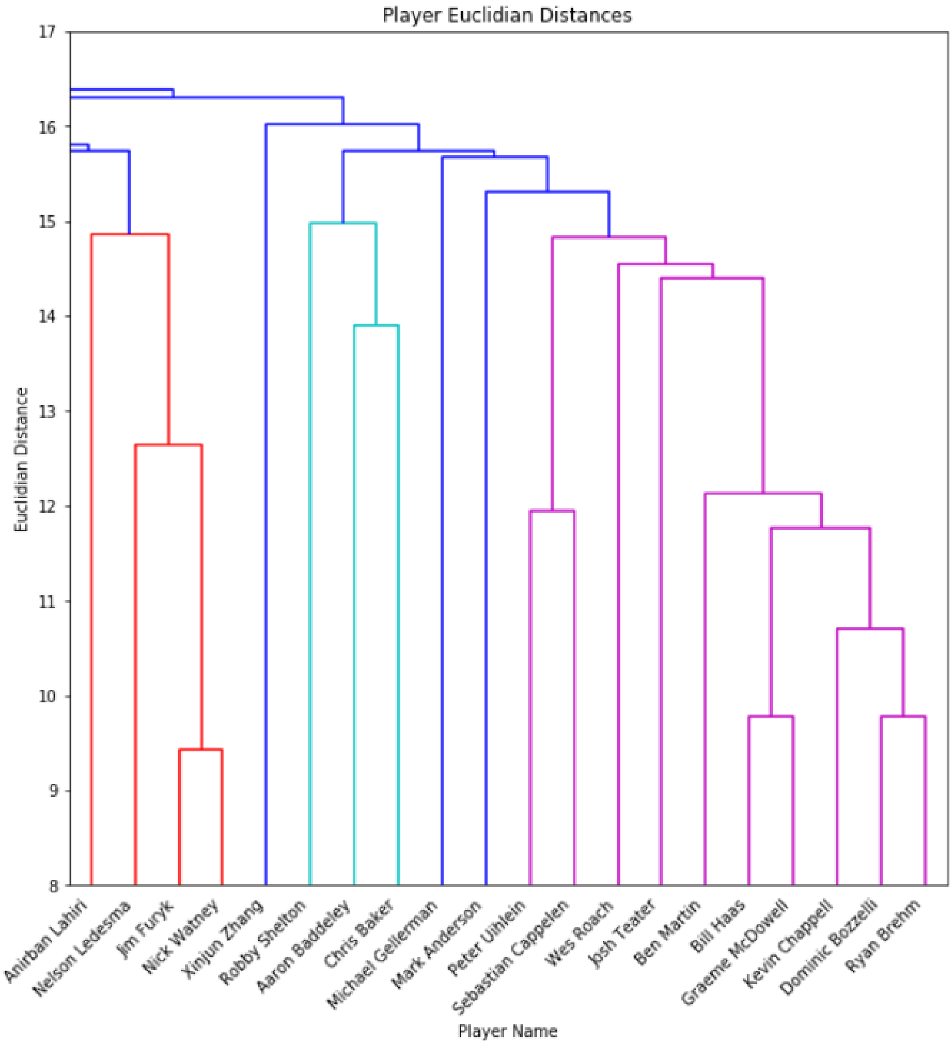
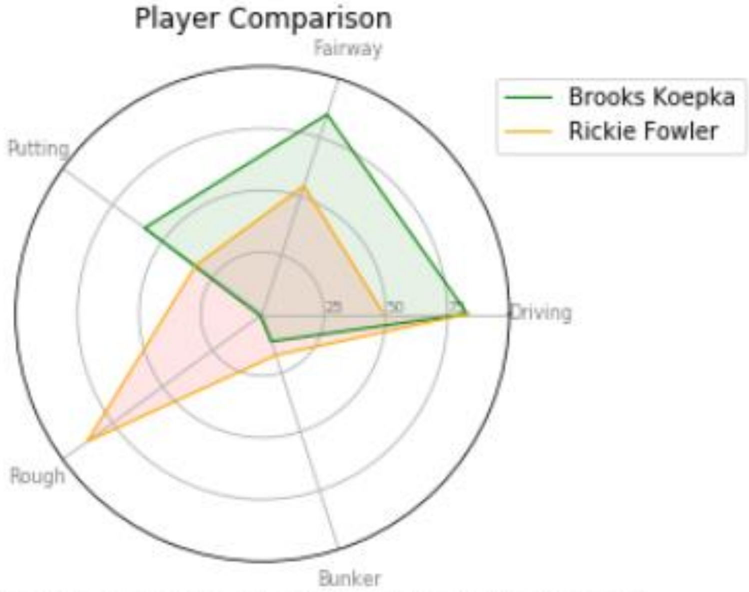


Figure 5: Visualization of a Euclidian Distance Matrix between Players

In *Figure 5*, the 3 colors represent groups of players with similar results across all tournaments. If we look at the lowest split, Jim Furyk and Nick Watney appear have the closest relationship between results. The full set of matrices and resulting dendrograms can be seen in the exploratory analysis workbook.

A more common player comparison visual we see is the spider plot. The spider plot takes different elements of a player's stats and plots them all in the same chart. If we overlap the spider plots of two players, we can see in which areas they perform similarly, and in which areas they differ. Let's look at the following example with Brooks Koepka and Ricky Fowler.



*Putting, Driving, and Fairway from Strokes Gained. Rough and Bunker from Green Scramble %

Figure 6: Spider Plot Comparison of Brooks Koepka and Ricky Fowler

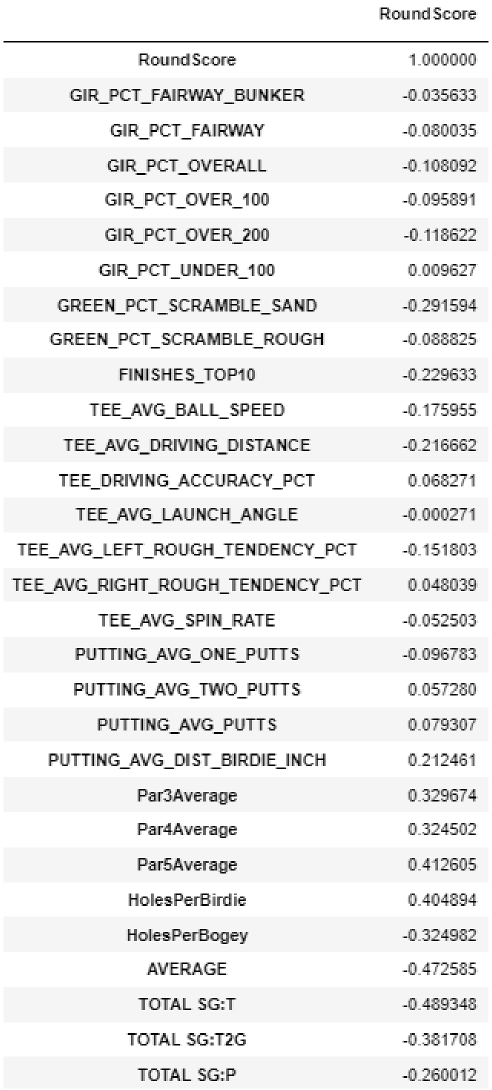
The spider plot in *Figure 6* uses Strokes Gained and Greens in Regulation stats to calculate percentile ranks in 5 key areas of the game. From this comparison, we can easily see Koepka performs better in putting, driving, and from the fairway, while fowler excels in play from the rough and slightly outplays Koepka in play from the bunker. The spider plot is an aesthetic way to perform these kinds of stat comparisons across multiple players.

Regression

The linear regression model is by far the most exciting but difficult section of this analysis. Other sport regression models I have researched average R-squared statistics of .6-.7, suggesting it is very hard to derive a somewhat accurate model. Still, we give it an honest shot in this analysis and finish with some encouraging results.

The first step in our analysis is selecting dependent and independent variables for our model. For the independent variables, we should perform some feature importance calculations to see which variable may be influencing our dependent variable. For the dependent variable, I created models with 3 different calculations for performance. These three metrics were round scores across all tournaments, tournament scores all season, and individual tournament score. While the first two dependent variables were selected to increase sample size, they were also the hardest to predict, scoring extremely poorly in all iterations of the models I built. For this report, we will focus on the 3rd dependent variable, tournament results for a single tournament. Since I knew we would have a limited sample size, I selected the tournament with the largest sample of players from our database, which was the Players Championship. With only one season worth of data to work the sample size was limited to 132 players and results, which would argue against the validity of any model we build.

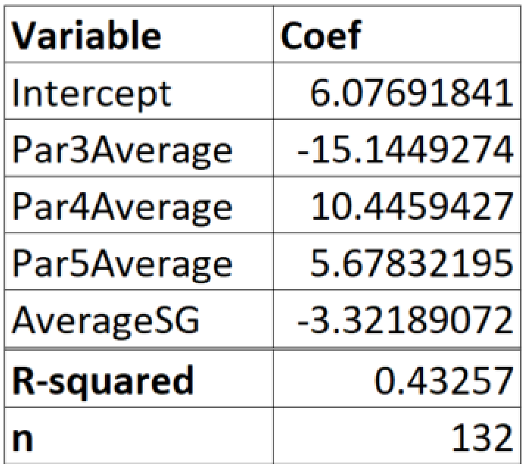
Continuing into feature selection to decide our independent variables, we created a correlation matrix between our dependent variable and player statistics. The results of this correlation matrix are shown below in *Table 2*.



	RoundScore
RoundScore	1.000000
GIR_PCT_FAIRWAY_BUNKER	-0.035633
GIR_PCT_FAIRWAY	-0.080035
GIR_PCT_OVERALL	-0.108092
GIR_PCT_OVER_100	-0.095891
GIR_PCT_OVER_200	-0.118622
GIR_PCT_UNDER_100	0.009627
GREEN_PCT_SCRAMBLE_SAND	-0.291594
GREEN_PCT_SCRAMBLE_ROUGH	-0.088825
FINISHES_TOP10	-0.229633
TEE_AVG_BALL_SPEED	-0.175955
TEE_AVG_DRIVING_DISTANCE	-0.216662
TEE_DRIVING_ACCURACY_PCT	0.068271
TEE_AVG_LAUNCH_ANGLE	-0.000271
TEE_AVG_LEFT_ROUGH_TENDENCY_PCT	-0.151803
TEE_AVG_RIGHT_ROUGH_TENDENCY_PCT	0.048039
TEE_AVG_SPIN_RATE	-0.052503
PUTTING_AVG_ONE_PUTTS	-0.096783
PUTTING_AVG_TWO_PUTTS	0.057280
PUTTING_AVG_PUTTS	0.079307
PUTTING_AVG_DIST_BIRDIE_INCH	0.212461
Par3Average	0.329674
Par4Average	0.324502
Par5Average	0.412605
HolesPerBirdie	0.404894
HolesPerBogey	-0.324982
AVERAGE	-0.472585
TOTAL SG:T	-0.489348
TOTAL SG:T2G	-0.381708
TOTAL SG:P	-0.260012

Table 2: Correlation Matrix with Tournament Scores

Using the correlation matrix as a starting point, a few iterations of the regression model were tried until achieving the best results using the independent variables: Par 4 Average, Par 5 Average, Par 3 Average, Average Strokes gained. I should also note I was able to achieve some higher scores with more explanatory variables but was hesitant to use them given the small sample size of the data. The variables were also min-max scaled so we could compare the coefficients and understand how each of the variables affects the results. The results of this regression model are shown below.



Variable	Coef
Intercept	6.07691841
Par3Average	-15.1449274
Par4Average	10.4459427
Par5Average	5.67832195
AverageSG	-3.32189072
R-squared	0.43257
n	132

Table 3: Regression Results

Results

The results of our regression model are shown in *Table 3*. Our final model achieved an R-squared score is 0.43 using the variables Par 3 Average, Par 4 Average, Par 5 Average, and Average SG. The coefficients can be confusing in the way they affect overall score, but the results can be read as having good stats in par 4 average, par 5 average, and average strokes gained will lead to a better score in the Players Championship. Unfortunately, this offers no insights on individual player skills that are useful for this tournament. Instead, it suggests that overall player quality are the most important stats for player performance.

Conclusion

To summarize the project, I was able to perform some interesting data transformations and analyses about the PGA Tour and its players. The exploratory analysis gives some interesting insights on the player based on their stats. Our player comparisons use both out player performance data and player statistic to compare strength and results. Although the regression analysis fell short, its R-squared score is not that far off other regression models used in sports betting. It would be interesting to see if the gap could be closed using data across multiple years for the same tournament, or player statistics from all players that participated in the tournament. Hopefully this would also allow the inclusion of some trait specific statistics that could help differentiate between golfers that will perform well across different tournaments.

References

Adam Maszczyk, Artur Gołaś, Przemysław Pietraszewski, Robert Roczniok, Adam Zając, Arkadiusz Stanula, Application of Neural and Regression Models in Sports Results Prediction, Procedia - Social and Behavioral Sciences, Volume 117, 2014, Pages 482-487, ISSN 1877-0428, https://doi.org/10.1016/j.sbspro.2014.02.249.(https://www.sciencedirect.com/science/article/pii/S1877042814017790)

Agarwal, Rahul. "The 5 Feature Selection Algorithms Every Data Scientist Should Know." *Medium*, Towards Data Science, 11 Sept. 2020, https://towardsdatascience.com/the-5-feature-selection-algorithms-every-data-scientist-need-to-know-3a6b566efd2.

Arastey, Guillermo Martinez. "The Increasing Presence of Data Analytics in Golf." *Sport Performance Analysis*, Sport Performance Analysis, 23 Jan. 2020, https://www.sportperformanceanalysis.com/article/increasing-presence-of-data-analytics-in-golf.

Keaveney , Bob. "How the PGA Tour Uses Data Analytics to Drive Powerful Fan Experiences." *Technology Solutions That Drive Business*, 4 May 2021, https://biztechmagazine.com/article/2020/07/how-pga-tour-uses-data-analytics-drive-powerful-fan-experiences.

Russo, Steve. "PGA Tour Stats - 2020 Season." *Kaggle*, 24 Jan. 2021, https://www.kaggle.com/steverusso/pga-tour-stats-2020-season.