Distance and Precision Affects Golf Player's Scores in PGA Tournaments*

Zhiyi Liu

27 April 2022

Abstract

Golf is about getting the ball into 18 holes with the least strokes. Players play competitively in world wide competitions such as the PGA tournament. The paper conducts linear regression model on PGA players' performances in each game to explore the factors that impact their scores. Knowing the relationships between strokes and distance, average putts and many more key performance metrics in golf is important for professional players to predict and improve. Keywords:Sports, Golf, PGA, Golf strokes

Contents

*Code and data supporting this analysis is available at: https://github.com/zzzhiyiliu/PGA_Scores.git

Introduction

The sport golf came from the Dutch word 'kolf', which means 'club'. Playing golf involves hitting balls into series of holes on a course in the least strokes. I am interested in exploring the key factors that drives performance because I want to understand how I can improve my scores. Given that golfing season is fairly short in Canada, I can only focus on a few points to practice because of the time constraints. I need to know where I should focus the most to optimize the effect of my practicing. Many follows the world wide golf competition for entertainment. Professional golf players compete in world tournaments such as the PGA. To explore the key variables that drives golf player's performance and their prize in the PGA, the paper uses statistical models to predict scores and earnings. Furthermore, the correlation between score and performance indicators are also examined.

The paper is divided into 4 sections including data, model, results and discussion. The dataset used in the study was collected from 2,312 rounds played by PGA professional golfers (PGA 2018). The data was provided on the PGA Tour website that includes number of rounds, strokes, average distance of strokes and other information on players' performance and earnings for that game from 2010-2018. All variables are numerical except for the names of the players. Since professional players play very consistently and well, the statistics are fairly close to each other.

This paper explores the factors that are correlated with the overall strokes of the games by multiple linear regression model. With all significant factors that might affect the performance, a linear regression model will be constructed to find the model that can best represent the score. The score is also paired with the number of wins in the year and see how they are correlated.

The results show that all factors are very significant in predicting the score. Only average putts is positively related with the score, higher fairway percentages, green in regulation, and average distance contribute to lower score. The score is also negatively correlated with the number of wins a player has in the year. The correlation coefficient is -0.478. Indicating that players usually wins tournaments will also very likely to perform well in PGA.

Data

Source

The data were collected from PGA tournaments for golf professional players. It contains all data recorded during 2010 to 2018. One observation represents one game played by one player in the PGA. One player might played multiple games in one year and for multiple years during 2010-2018. The dataset is available for download from the PGA website.

Methodology

The dataset used in the study was collected from 2,312 rounds played by PGA professional golfers. Variables about the players such as rounds played in the PGA, number of times in the top 10 and number of wins are collected based on the player's information. Other variables are collected based on the performance of that round, such as fairway percentages and average putts. Information was collected for all holes played. One exception being that the average distance is only measured from one hole that players are likely to hit drivers off the tee, the first shot of a hole, in the PGA tournaments (CDW 2016).

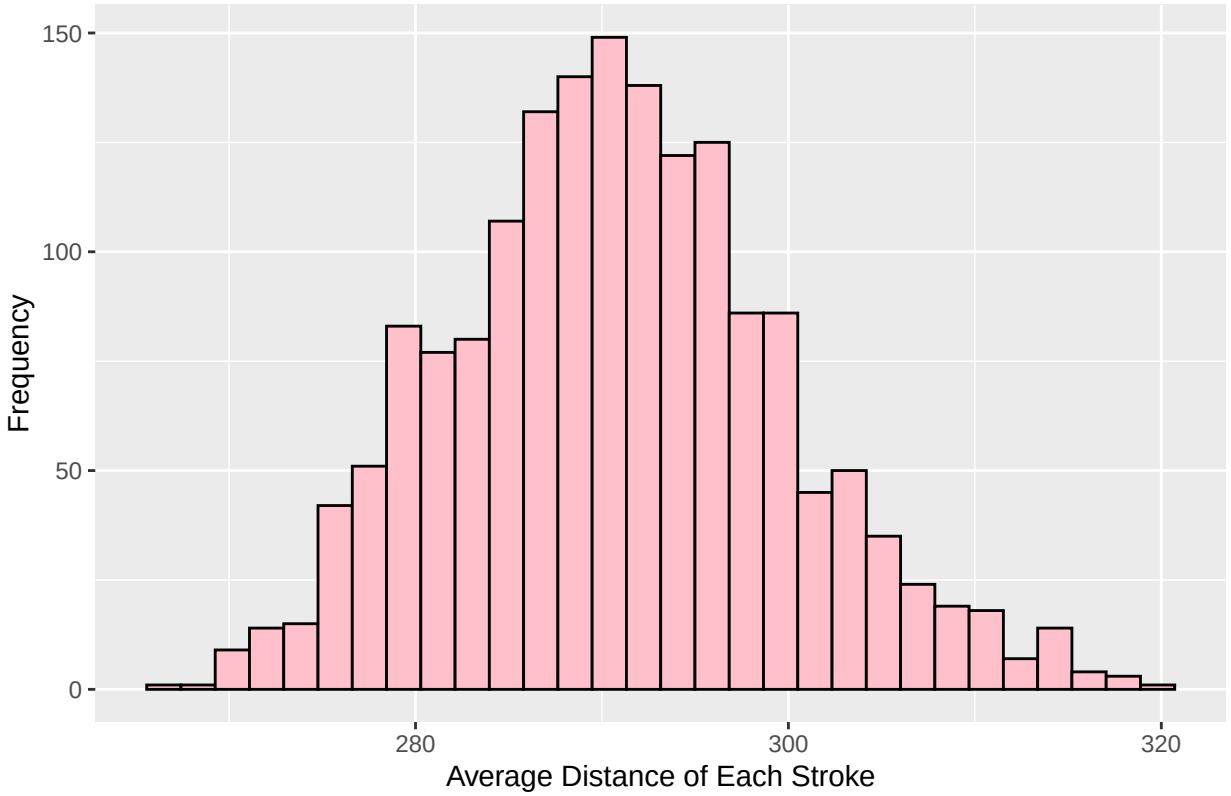
The data is cleaned and the model will be analyzed through R studio, an open sourced statistical anlayzing program (R Core Team 2020). Packages used in the paper includes tidyverse and tinytex (Wickham et al. 2019) (Xie 2022) . Plots are generated through ggplot (Wickham 2016). The report was generated through knitr (Franbois, Henry, and Miller 2021).

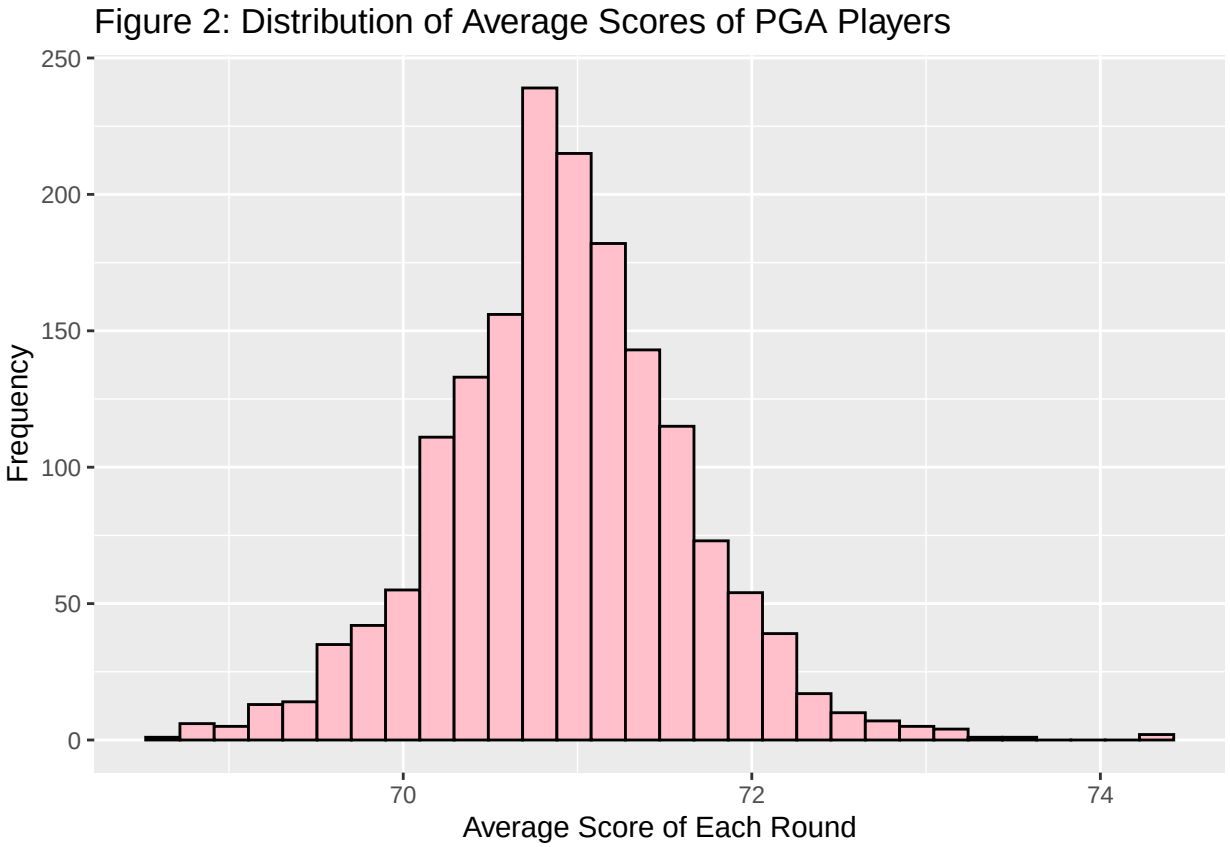
Data Characteristics

There are 18 variables in the dataset. All variables and terms are explained in the appendix section (Appendix 1). It contains information about the general performance of the players. There are also 5 variables that are focused on the statistics about players' short game performances. Short game starts when the ball is near the green or on the green. A chip or a putt would be considered as short game. All variables are in numerical form. Only the year number, number time in the top 10, points and number of wins are discrete numerical variables.

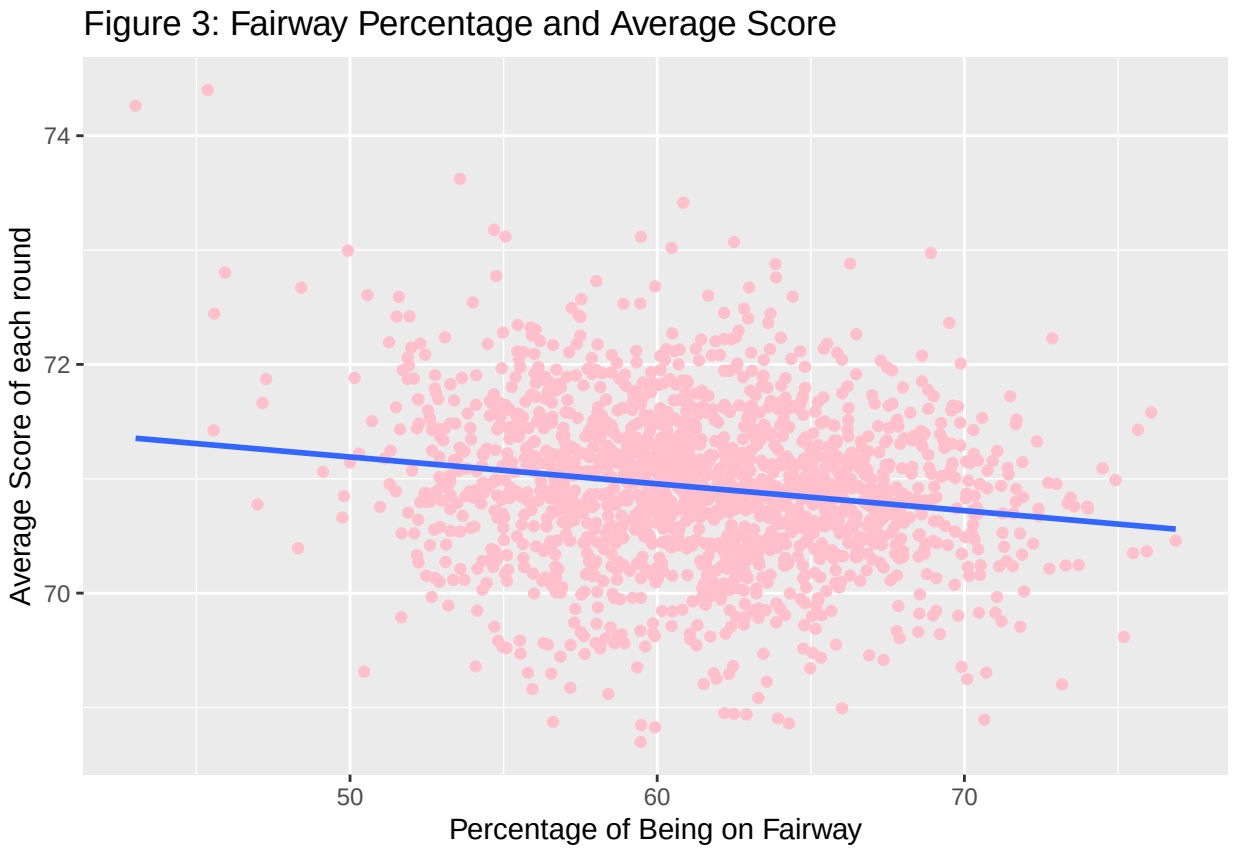
The dataset had 1678 observations after removing missing datapoints. Almost all variables are normally distributed. The spreads of all variables are small since PGA players perform consistently. Since the datapoints has too much proximity, it might be difficult to see the effect of predictors on the response variable.

Figure 1: Distribution of Average Distance of PGA Players



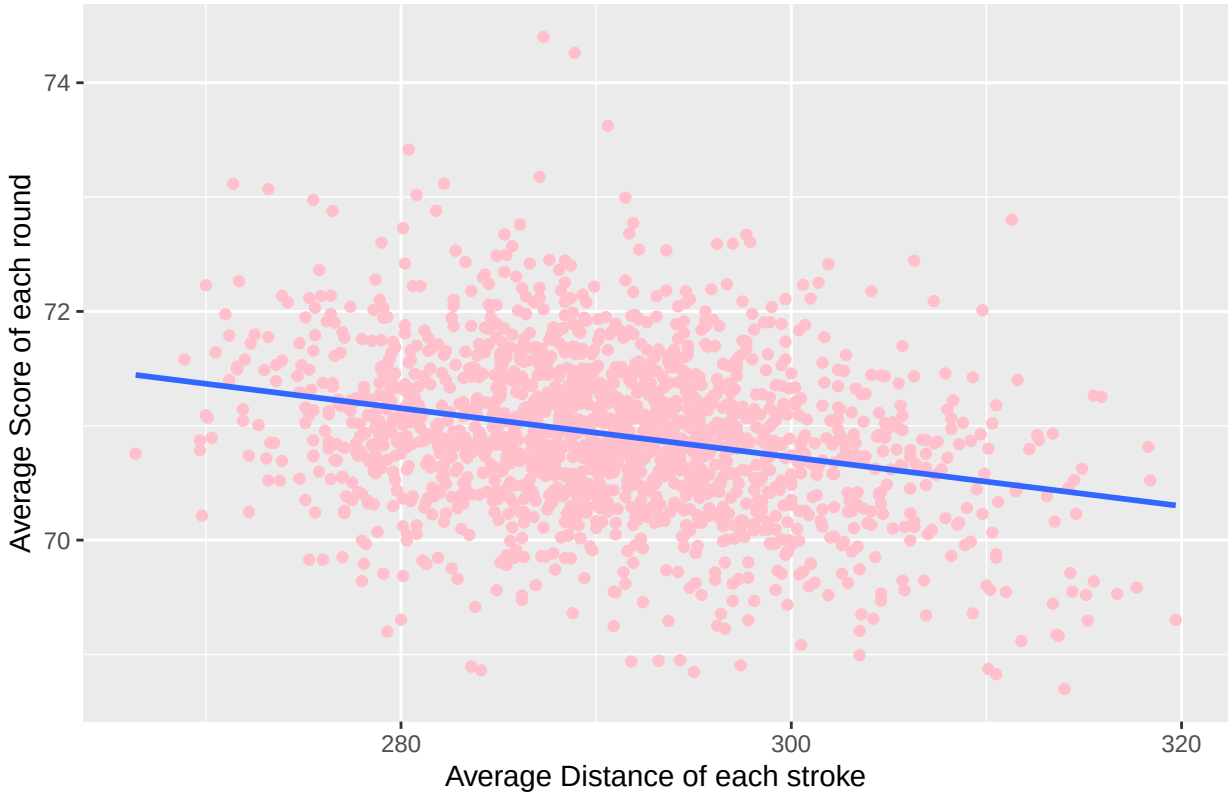


The mean distance of PGA players is 290.8 yards (figure 1). The median is 290.6, meaning that there are equal players that have shorter distance than longer distances. The standard deviation for distance is very small, only 8.9 yards. The average score of the rounds is 70.92 (figure 2). The standard deviation is only less than 1, 0.698. This means that professional players perform consistently in playing tournaments. The range of scores recorded only has a difference of 5.7 strokes.



The fairway percentage and the score has a slightly negative relationship, as seen in figure 3. Both variables are fairly equally distributed along the regression line. Figure 4 shows that the average distance has a slightly more slanted slope in the regression line with the average score. Both variabels have a negative relationship with the score.

Figure 4: Average Distance and Average Score



Model

The model used to determine the factors that influence the score of the round is multiple linear regression model (MLR). The initial model's predictors are chosen based on intuition. Average distance of shots, fairway percentages, average putts and green in regulations are chosen as predictors in the initial model. The response variable would be the score of the round. The purpose is to see if one of those variables is more impactful on the overall score.

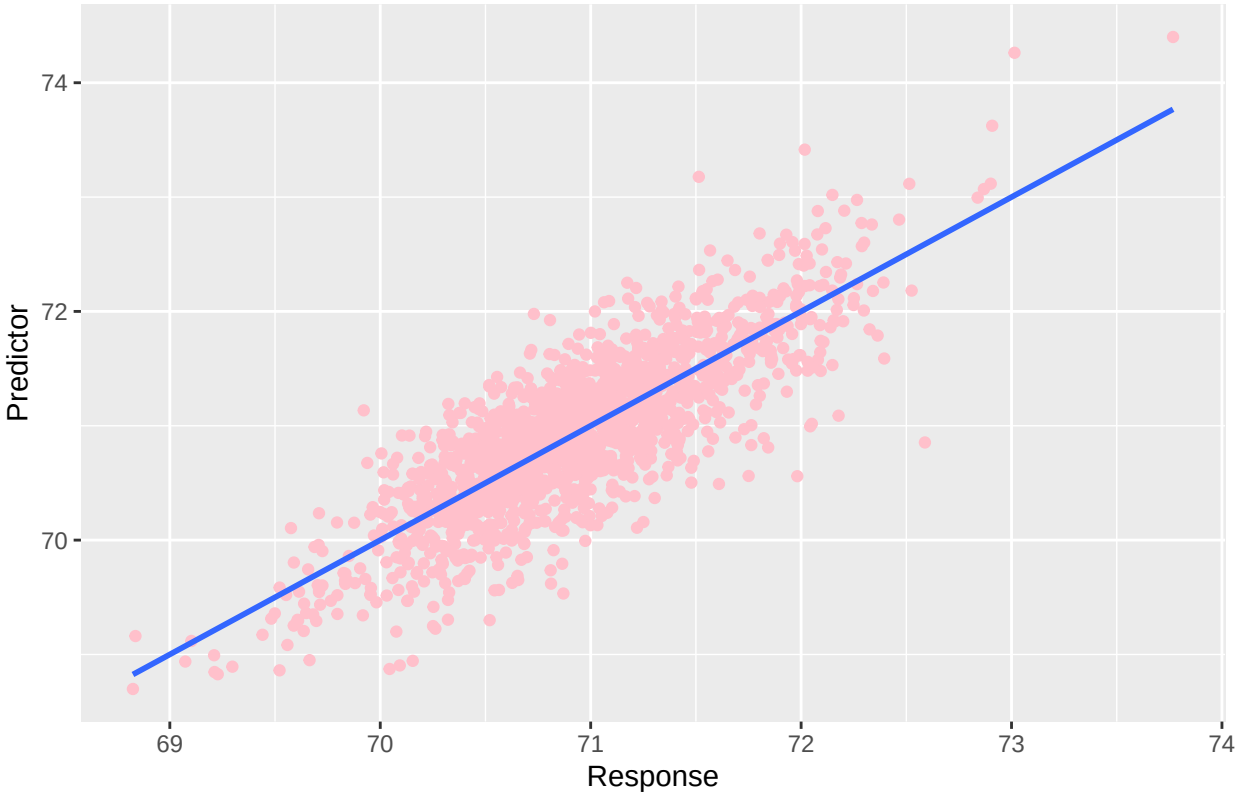
Regression Model on Score

Before conducting the linear regression model, we have to check the assumptions of linearity, independence and constant variance. The linear model to predict the score of each game can be conducted by starting variables based on assumptions. Potential predictors that might influence strokes include average distance of shots, fairway percentages, average putts and green in regulations. The model will be examined by the partial-f test to see excessive predictors to our response variable.

Condition 1

The initial model is constructed by the 4 variables mentioned above. The model is being tested on the assumptions. First of all, the conditional mean response is a single function of a linear combination of predictors.

Figure 5: Response and Predictor



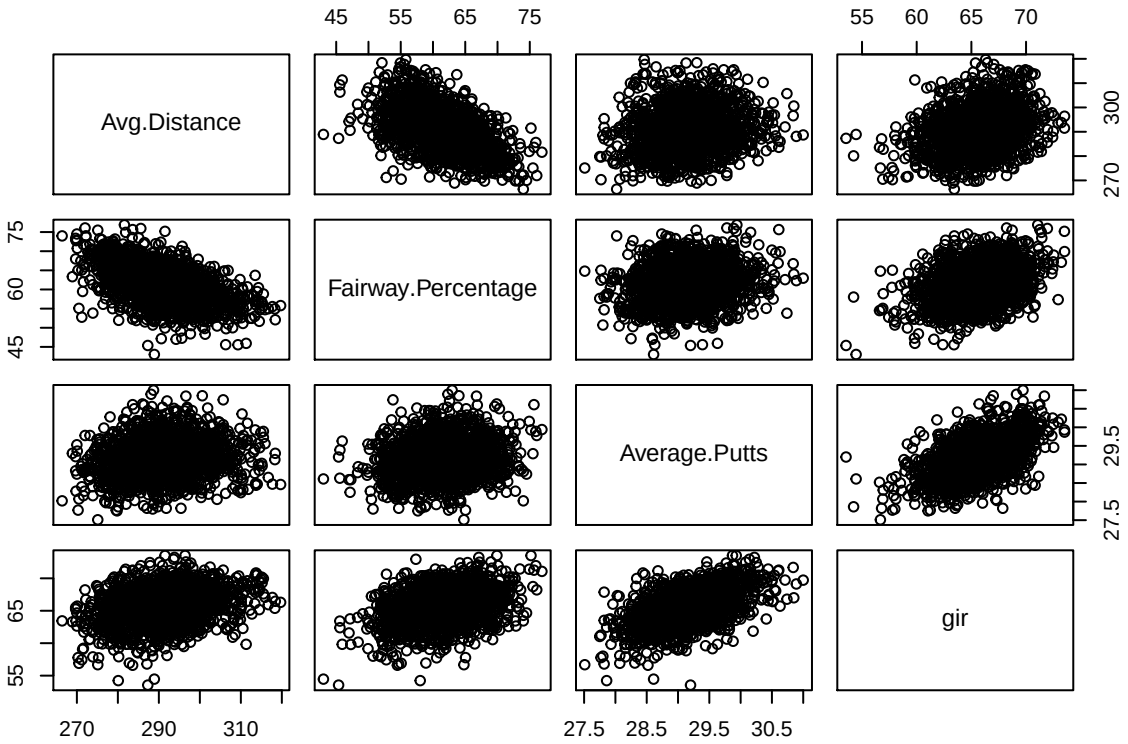
The condition 1 is satisfied. In figure 5, it shows that they have a linear relationship, the regression line is very clear.

Condition 2

The second condition a multiple linear regression model is that the conditional mean of each predictor is a linear function with another predictor. The predictors either must have no relationship or a linear relationship with each other.

The condition can be examined when the scatterplots are presented with the relationship between each predictors. As we can see through figure 6, all relationships between the predictors satisfy the conditions.

Figure 6:

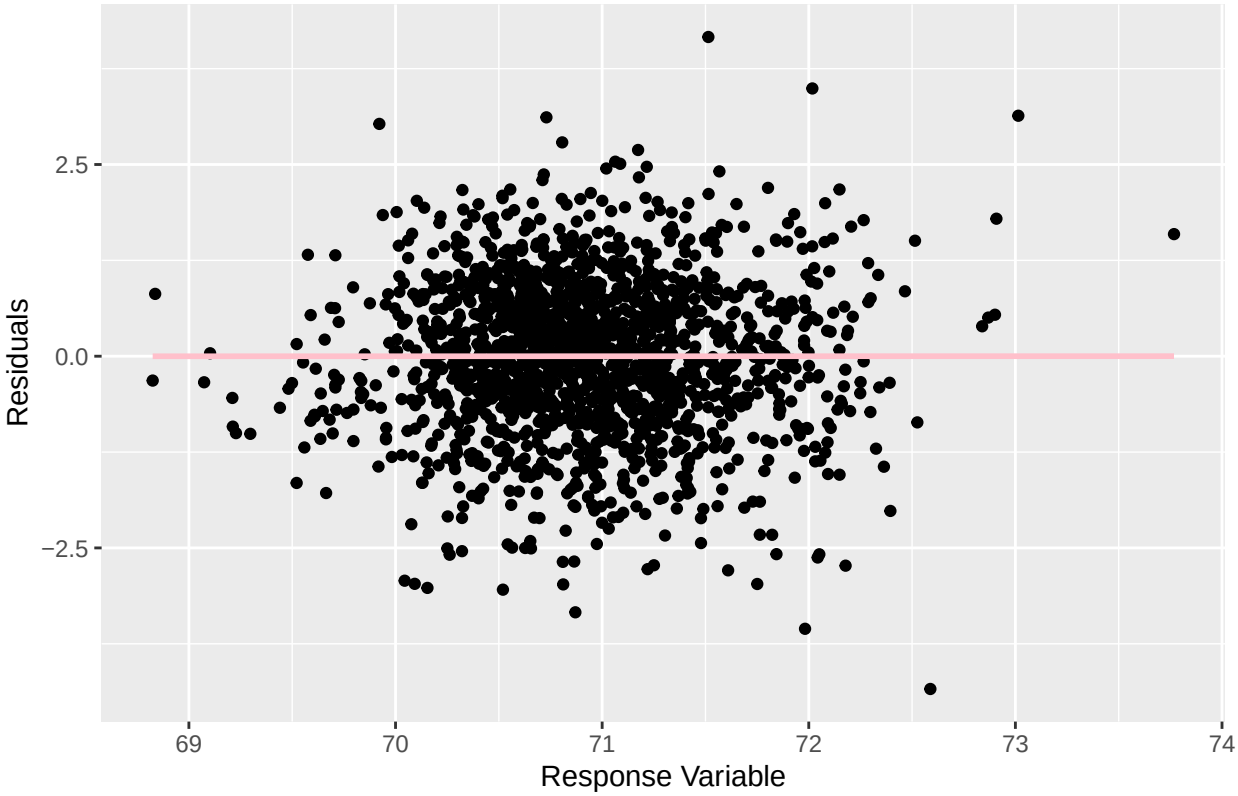


Residual and Fitted Model

The residual plot can help us to determine if the model violates any assumptions on linearity, normality and constant variance.

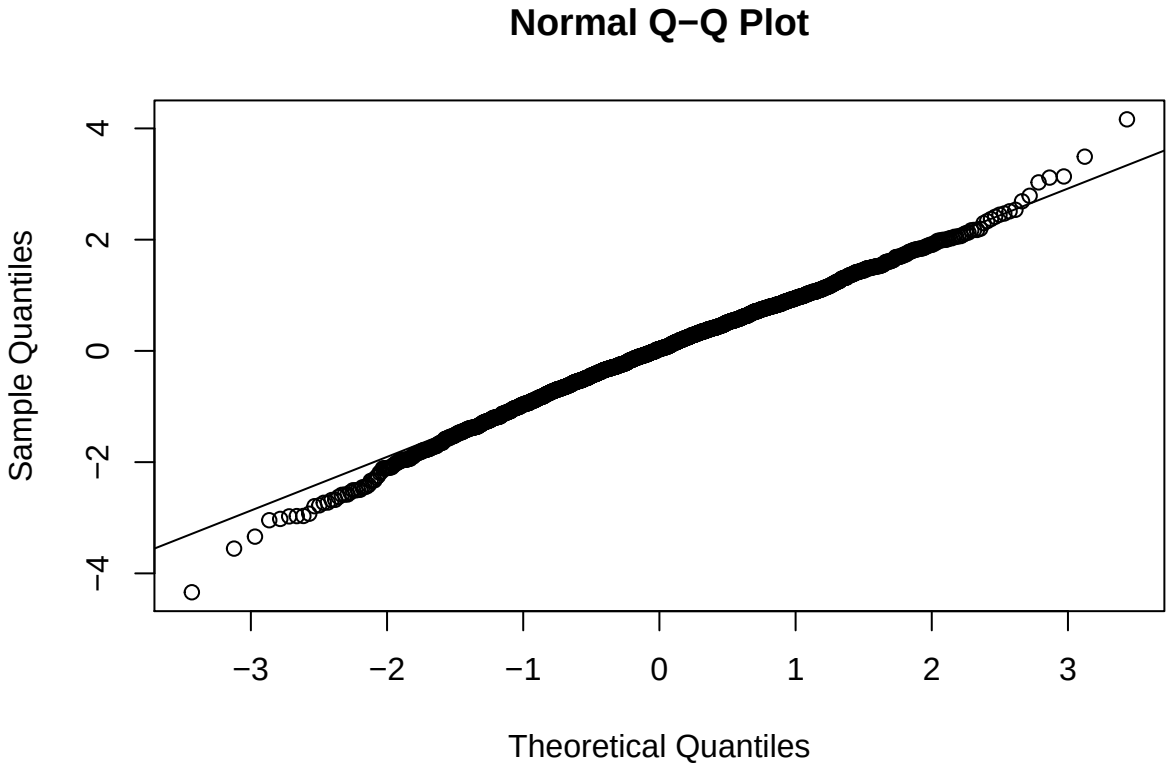
The scattered plot below shows the residual and the fitted value. The graph shows that points are mostly clustered in the center around 70.5 and 71.5. Depending on the interpretation, the independence assumption would be considered satisfied. However, some may argue that the independence assumption could be further examined in this case. For the small standard deviation of the variables itself in the original dataset, we can say that the cluster is negligible. The values are equally distributed along the regression line with the residual of 0. We can say that the linearity is satisfied. The constant variance assumption is also satisfied.

Figure 7: Response and Residuals



The normality assumption for MLR is also not violated. The points are all along the 45 degree line. The points on the end are slightly deviated from the line itself. However, it is still considered normal given that the deviation does not impact the normality. We can conclude that all assumptions to conduct the MLR are satisfied (figure 8).

Figure 8:



The number of competition a player has won is also examined. The simple linear regression model will be conducted to see if there is any correlation between the number of wins and the performance of the players. The purpose is to discover a particular trait of the players can bring consistent performance in their career. For instance, player who has higher fairway percentages might not have good score in that round, but might have outperformed the other players throughout the season.

Results

Performance Factors

The multiple linear regression model was conducted to determine the model that can best estimate the scores by using average distance, fairway percentages, average putts and green in regulation.

The result shows that the intercept of the linear model is at 61.663. The betas for average distance, fairway percentages, average putts and green in regulation are -0.21293, -0.25508, 1.016044 and -0.193032 respectively. This indicates that only average putts contribute to a positive number of score. The other variables are negatively contributing to the number of stokes in a round.

Figure 9: Mutiple Linear Regression Model Result

	Estimate	Standard Error	t-value
Intercept	61.663	0.701	87.91
Average Distance	-0.021293	0.001622	-13.120
Fairway Percentages	-0.25508	0.002913	-8.412
Average Putts	1.016044	0.022622	44.914
Green in Regulation	-0.193032	0.005340	-36.148

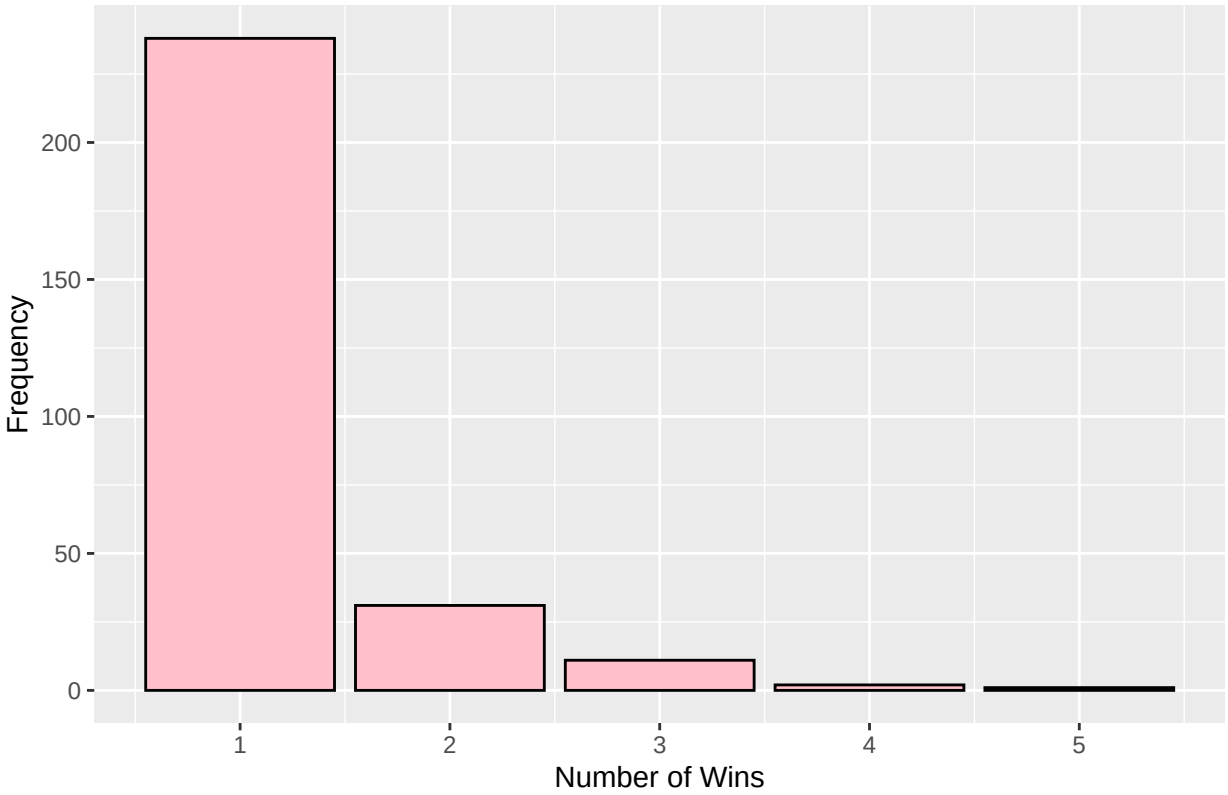
All p-value of the variables in the model are very small, smaller than 2e-16. Meaning that all variables are equally important in the model in terms of contributing to predicting the score of the round (figure 9).

The residual error is 0.401 with 1673 degree of freedom. The residual error is relatively small.

Score and Number of Wins

The paper further examines if the score is correlated with the number of times the player has won a competition in that year. Intuitively, the ranking of players in each game in PGA would be determined by their score. We want to see if there is a relationship between the previous wins a player has and the current score of the round they plays.

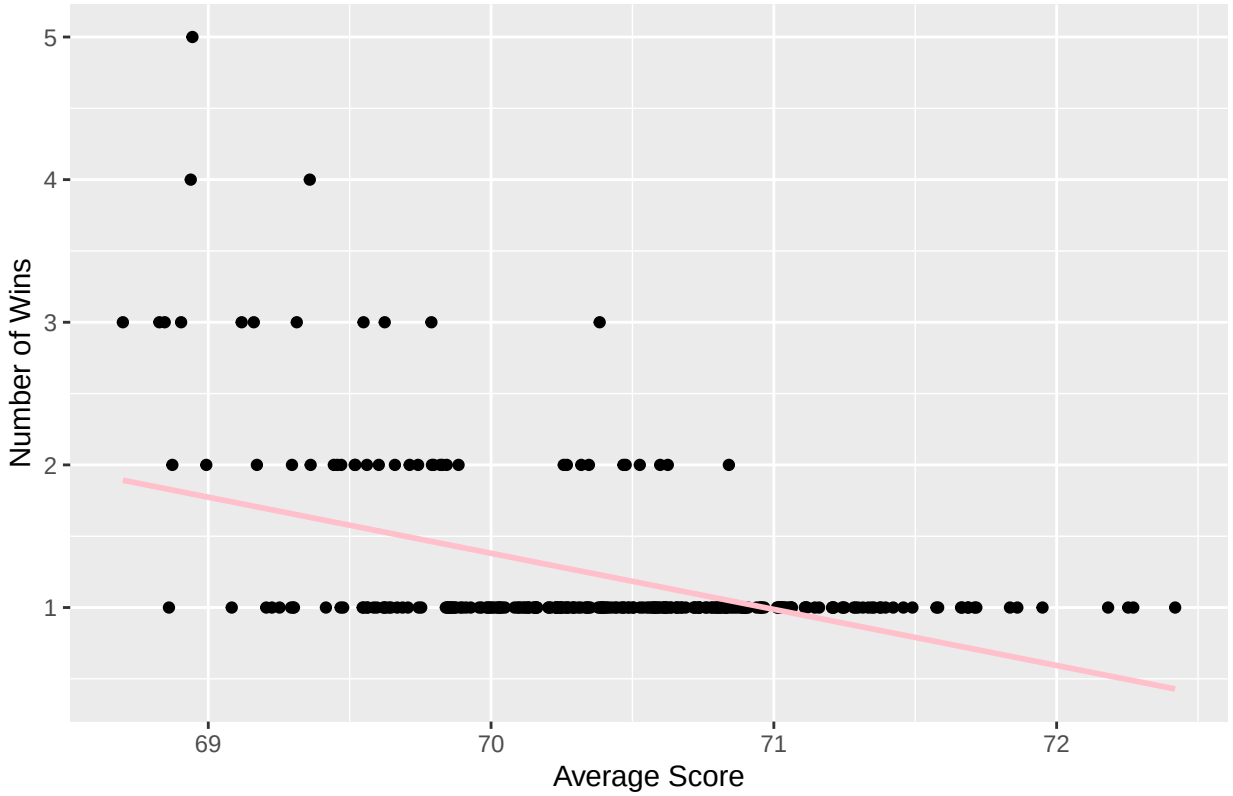
Figure 10: Number of Wins of PGA Players



We can see that almost 250 out of 1678 observations have had 1 win in the year (figure 10). The distribution is right skewed. There are fewer people who has more wins. The maximum is 5 wins in the year. The statistics also indicate that there are more than 1400 rounds that player played with no wins at all in the year. That is more than half of the rounds recorded in the dataset. We can say that most players did not win a single game in the year.

Furthermore, we examine the relationship between the performance of the round and the previous wins by the simple linear regression.

Figure 11: Players Average Scores and Number of Wins in the Year



Since most people does not have a win in the year, we only examined those who did won at least one competition in the same year. The plot above is only conducted based on 283 observations. We can see that the regression line has a negative slope. The lower the score is, the higher number of wins a player will get (figure 11).

After conducting the pearson's correlation test, we found that the correlation coefficient between number of wins and scores is -0.478. This indicate that the two variables have a moderate negative relation. The 95 percent confidence intervals is between -0.563 and -0.382. The more wins a player has in the year, the better score they are likely to get in PGA tournaments.

Discussion

The purpose of the paper is to explore what factors that can influence golf scores the most. The data downloaded from the PGA tournaments from professional players have suggested that many factors are significant predictors for the scores. The scores can be expressed as a linear model with multiple variables including average distance, average putts, green in regulation and fairway percentages. The number of wins a player has in the year can also be a good indicator of how the player is going to score in the PGA. The relationship is moderate, reflecting that winners are likely to score lower in PGA. The residual error is 0.4. Even though it is not a big number, perhaps adding more variables in the initial model can provide a better fitted model to predict the scores of PGA players

Average Putts

The final linear regression model had a intercept of 61.663. The beta for the average putts as the predictor of the response variable average score is 1.01644, the only positive beta for the model. The result makes sense since the final score includes all strokes of a player during a round. Both a long-shot and a putt would count towards one shot. Even though a player might be better at short-games (chipping and putting), indicating a lower average putt overall, 1 putt counts towards 1 score. Subsequently, the score of the player increase by 1.

The beta reflected the reasoning. If one player accidentally made a mistake in one of the holes and had an extra putt, the final score would also increase 1.

The average putts is also one of the most independent predictors among all four variables used in the model. All other variables are focused more on long-games, as the average putt is the only variable that captures the short-game of the players' performances. Players might have more competitive advantage if he has a strong short game, giving lower average putts per game. Since there is only one variable, the large beta makes sense in this case.

Moreover, professional players tend to put more emphasis on short games since long-games (drivers, woods and irons) should be relatively stable in tournaments. In professional tournaments like PGA, players are world wide talents who has both distance and precision in their long-games. As the results for the average distance showed, the standard deviation is less than 9 yards. The key to accelerate in the career is to focus on the short games.

For the original intention of the paper, to see what should I focus more to improve my score, the results gave a good idea of the importance of short game. In this case, putting directly affects the number of strokes. I should be focusing on practicing on my putting. However, we cannot ignore the the fact that PGA professional players have almost perfect golf swings that give them consistent performance on long-games. Focuses should also be put on driver, woods and irons in order to improve the score.

Distance and Precision

On the other hand, the better precision and distance contribute to a overall good score (lower strokes) also makes sense. Even though the betas are not as significant as the average putt predictor in terms of absolute value, they are also strong predictors of the performance.

If a par 4 or par 5 hole requires the player to drive long, the player who has further driving distance would have an advantage. After the tee-shot, where the PGA measures the average distance variable, the player who stroke the furthest would be closest to the hole, assuming that all players are able to manage to get the ball in the same direction towards the hole. It would also be intuitive to think that the difference between the shot distances across players would make a large difference on the performance accumulative in a tournament. Overtime player strokes the first shot from the tee, players with longer distance would have a few yards advantage. In the long-run, the yards would accumulate into distance that requires extra stroke to reach. However, the distance is only measured on the first shot off the tee. It does not capture what happens after the first shot.

For instance, a player might have a solid driver shot with 300 yards as average distance but performs below average in playing on the fairway. Other players might have a shorter driving distance but has better precision to get the ball on the green when it is on the fairway. Precision might be a more significant indicator for late game. This explains why the beta for fairway percentages, representing precision, is lower than the beta of the average distance. -0.25508 has a larger impact than -0.21293, as our model suggested.

The distance factor would be more applicable to longer holes, not in par 3 holes. The total distance of the hole must be longer than the player's driving distance in order to be significant in impacting the final score. Another scenario where having long distant shots can be beneficial is when there is an obstacle. Players cannot hit far might have to take a short shot to land before the obstacle and take another shot to pass it. People with long distance can pass it without the second stroke. That would contribute to a lower score with an advantage. However, this is only applicable to certain golf courses and specific holes, not very significant in predicting the overall performance in the tournament.

Green in regulation, subsequently is a indicator of both distance and precision. It indicates if the player get the ball on the green with 2 fewer strikes than the par. For instance, a player uses 1 shot to get on the green in a par 3 would be green in regulation. If a player does not get to the green on the 3rd shot of a par 5, the hole cannot be counted as green in regulation because the player would need a 4th shot to get the ball on the green.

The beta for the green in regulation is -0.19, the least significant among all predictors. Even if the player

gets the ball on the green with 2 shots less than par, the short game is so important that the long-game is not as significant. This further supports the claim that short-game is the key to improve performance for professional players.

Number of Wins

Whether winning in competitions is aligned with their performance in the PGA tournament was examined through the simple linear regression model. They are plotted as a scattered plot with a regression line in the middle, indicating that there is a clear trend in players who won more had lower scores. Most players who only had 1 win has an average score between 70 and 71. This indicate that, if you want to win more competitions, average score in the game must be below 70.

The data only showed the number of wins the player had only when it is above 1. If the player did not have any win, the dataset does not include a value in the wins variable. There might be a possibility that some players did not participate in large amount of competitions, thus winning none. Those players might have only came to the PGA. When the data was cleaned from removing empty cells, only 283 observations were left. The rest either did not participate or won none.

It would make sense to think that players who play well, who have lower average scores, would perform well in all competitions. The regression model also showed that it is correlated with the performance of players. People can then use the number of wins to predict the performance of PGA players in future games. However, the correlation coefficient suggest that it does not have a very strong relationship. It would be reasonable to use it as a predictor for general placements of the players but not accurate enough to tell the scores. Future studies can take longer data and simulate a model to predict performance based on historical data.

Limitations

One of the weaknesses of the paper is that the initial model only contained 4 variables and was selected based on intuition. As someone who has experience in golf, I chose the predictors based on my knowledge. There might be some bias when selecting the variables. If it was a large model with 10 predictors, it would make more sense to choose the variables on intuition. The big model can be reduced by conducting the partial F test to see if there are redundant predictors that can make the model better. In the case of this paper, the model only had 4 variables that is hard to do so. There is also the possibility that some factors that impact the scores are left out from the model.

The model can be perfected by see if adding more variables would yield a smaller error in predicting the response variable. Future projects can also conduct anova test to see which model is better. Furthermore, since the number of wins has a moderate relationship with the score, the model can also try to incorporate the factor to see if it can yield a more accurate model.

The purpose of the paper is to help identify key areas to focus for a non-professional golf player to improve skills. A fundamental problem is that the data used to do the analysis are based on the performance of PGA players. There might be a different focus in practicing for non-professional players like me. The results in the paper can be only used as a reference to the performance for professional players. It might be more useful in predicting the scores of PGA tournaments for future years. For the original intention of the paper, future projects can also examines how golf players improved their score and their performances as a result of their focuses. MLR on how many hours spent on each area (driver, iron, putting, core exercises, playing on course) can be conducted for future projects.

Appendix

Appendix 1:

Variables: Player Name: Name of the golfer

Fairway Percentage: The percentage of time a tee shot lands on the fairway

Avg Distance: The average distance of the tee-shot

gir: (Green in Regulation) is met if any part of the ball is touching the putting surface while the number of strokes taken is at least two fewer than par

Average Putts: The average number of strokes taken on the green

Average Score: Average Score is the average of all the scores a player has played in that year

Wins: The number of competition a player has won in that year

Appendix 2:

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

- The data was created to record the performance of PGA players every year.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

- The data was recorded by PGA officials who want to keep track of players performances.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
    - The dataset contains performance of the PGA players such as the score, average distances, fairway percentages, etc.
2. *How many instances are there in total (of each type, if appropriate)?*
    - There are 2312 observations, each representing a round played in the PGA.
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
    - The dataset contains all instances. The dataset is representative of the players who played in the PGA tournament.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
    - The dataset contains process data by the individual
5. *Is there a label or target associated with each instance? If so, please provide a description.*
    - No
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
    - No
7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
    - No
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
    - No
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
    - No, the only issue is that the dataset have each round as an observation instead of each player. A player might played multiple times that might impacted the result of the analysis.
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - It is self-contained.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected*

by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
   - No
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
   - No
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
   - The dataset was only recorded on PGA professional players.
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
   - No
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
   - No

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*
   - The data was collected based on 2312 rounds played in the PGA tournament. It was recorded after the games were played. The data was directly observable. The data are mostly averages of the performance a player did in the tournament. Some are based on the historical data of the player before the game was played, such as the number of wins and the number of times the player was in top 10.
2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
   - The PGA officials has to count all scores and the details in how the players performed. Ex. how many putts, whether if the ball in on green, etc. One exception being that the average distance is only measured from one hole that players are likely to hit drivers off the tee, the first shot of a hole, in the PGA tournaments.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
   - No
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
   - No
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
   - The data collection from 2010-2018
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.* -No
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
   - no
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with*

screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
   - No
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
   - The entire tournament is also recorded, players must consent that their data will be displayed to the public.
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
   - No
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
   - No

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
   - Some missing variables are removed at the beginning.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
   - N/A
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
   - N/A
4. *Any other comments?*
   - N/A

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
   - No
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
   - No
3. *What (other) tasks could the dataset be used for?*
   - The data can also be used to analyze individuals' performance over the past years.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
   - No
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
   - No

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
   - No

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
   - N/A
3. *When will the dataset be distributed?*
   - N/A
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
   - N/A
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
   - No
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
   - No

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - The statistics part of the official website is updated every year.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - they can be reached through https://www.pgatour.com/stats.html
3. *Is there an erratum? If so, please provide a link or other access point.*
   - No
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
   - N/A
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
   - No
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
   - No

Reference

CDW. 2016. *STATISTICALLY Speaking Statistically Speaking: Driving Distance Key?* https://www.golfchannel.com/video/explanation-pga-tour-driving-distance-statistic.

Franbois, Yihui Xiein, Lionel Henry, and Kirill Miller. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in R.*

PGA. 2018. *Statistics.* https://www.pgatour.com/stats.html.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2022. *Tinytex: Helper Functions to Install and Maintain Tex Live, and Compile Latex Documents.* https://github.com/yihui/tinytex.