**Project Report**

**PGA Tour Stats**

**1. Introduction**

Golf is a sport that blends skill, strategy, and consistency, with success often determined by a player’s ability to excel in key performance metrics. The PGA Tour, one of the most prestigious professional golf tours, provides a fountain of statistical data that can help identify what separates the best golfers in the world from the rest of the field. By analyzing player performance from two distinct time periods—2010-2018 and 2022-2025—we aim to discover patterns and key factors that contribute to success on the PGA Tour.

Professional golf has evolved significantly over the past decade, with advancements in technology, fitness, and course strategy influencing player performance. Metrics such as driving distance, greens in regulation, strokes gained, and putting accuracy have become critical indicators of success. Modern analytics like "Strokes Gained" have reshaped how players and analysts look at performance. “Strokes Gained is a statistic that compares a golfer’s performance to other golfers. Specifically, Strokes Gained measures the number of strokes a golfer gains or loses relative to the average” (golfshot.com). By using data from both time periods, we can compare how these factors have changed over time and determine which statistics most reliably predict success in professional golf.

Our analysis will focus on identifying the statistical trends that are noticeable in elite golfers, examining whether certain skills have become more or less important over time. This will provide valuable insights into how players can optimize their game to compete at the highest level in today’s golf landscape.

**2. Data**

To analyze professional golf performance and find patterns among PGA Tour players from 2010 to 2018 and 2022 to 2025, we utilized two primary data sources: a publicly available dataset from Kaggle and statistics obtained via web scraping from the official PGA Tour website. These two data sources were merged to create a player statistic dataset that allowed us to explore performance metrics such as driving distance, putting accuracy, and scoring efficiency.

The first dataset came from Kaggle which was titled “PGA Tour Data 2010–2018.” This dataset includes an in-depth set of statistics for professional golfers on the PGA Tour, spanning a nine-year period. For each player-year combination, the dataset contains quantitative metrics such as fairway percentage (an indicator of driving accuracy), average driving distance, greens in regulation (GIR), average putts per round, scrambling percentage (ability to recover from missed greens), average score, and the number of FedEx Cup points earned. This dataset served as the backbone for our analysis.

To further enhance our dataset and continue to evaluate player statistics, we also incorporated live data scraped from the PGA Tour’s official website. Using Selenium, we extracted updated values for average driving distance and total rounds played per player. These values were matched against the Kaggle dataset by player name. In cases where the web-scraped data provided more recent or complete information, it was used to fill in missing or inconsistent fields from the original dataset.

After preparing both datasets, we continued working in Jupyter Notebook and used a left join on the “Player” field to merge the two sources. During the merge process we carefully reconciled duplicate columns, such as those for “Rounds” and “Avg Distance”, by selecting the most complete or recent values available. Rows containing missing values for statistics like “Fairway Percentage” or “Average Putts” were dropped to maintain the quality of the data. The result was a unified dataset containing reliable performance information for PGA Tour players.

The final merged dataset included these columns:

* Player: The name of the professional golfer.
* Year: The golf season year associated with the player’s performance.
* Fairway Percentage: The percentage of times the player successfully hit the fairway off the tee.
* Avg Distance: The average driving distance in yards, reflecting power off the tee.
* GIR: Greens in Regulation percentage, indicating how often a player reaches the green in the regulation number of strokes (Par minus two).
* Average Putts: The average number of putts taken per round, representing putting efficiency.
* Average Scrambling: The percentage of successful recoveries after missing the green in regulation.
* Average Score: The mean score per round across events played in a given year.
* Points: The total number of FedEx Cup points earned in a season, often reflecting overall performance and consistency.
* Performance Group: A derived categorical variable that classifies players into “High” or “Low” performance groups based on a median split of FedEx Cup points.

By combining historical statistics with current web-scraped performance data, we created a dataset capable of supporting many different analytical techniques including regression modeling, hypothesis testing, and classification. This helped our group better understand what statistical attributes are most strongly associated with elite performance in professional golf.

**3. Analysis**

*3.1 Impact of Driving Accuracy on Overall Performance*

The first central analysis question we explored in this project was: Does a player’s driving accuracy significantly impact their overall performance on the PGA Tour? Driving accuracy, known in our dataset as Fairway Percentage, reflects how often a golfer successfully lands their ball on the fairway from the tee on par-4 and par-5 holes. Overall performance was measured using FedEx Cup Points, a standard metric that accounts for tournament results throughout the PGA Tour season.

We started by dividing players into two groups based on driving accuracy: those who hit more than 65% of fairways (“Accurate”) and those who hit 65% or fewer (“Less Accurate”). A box plot (Figure 1) was generated to visually compare the distribution of Points across these two groups. The plot showed that the “Accurate” group tended to have a higher median number of Points and fewer low outliers, indicating more consistent performance.

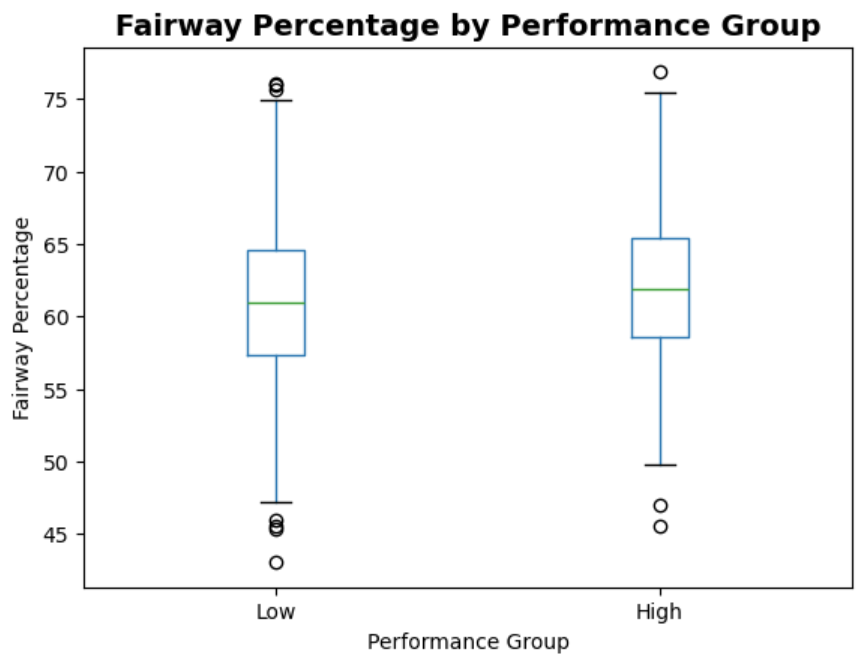


Figure 1: Boxplot of Fairway Percentage by Performance Group

A two-sample t-test was conducted to determine whether the difference in mean Points between the two groups was statistically significant. The test produced a t-statistic of 2.53 and a p-value of 0.011, indicating that the difference in performance between more accurate and less accurate players is statistically significant at the 5% level. This result supports the hypothesis that accuracy off the tee positively influences a player's season-long performance.

To further quantify the relationship, a simple linear regression was performed using Fairway Percentage as the predictor and Points as the response variable. The model produced an R-squared value of 0.076, suggesting that while driving accuracy does have a positive impact, it explains only a small portion of the variance in overall performance. This makes sense, as success on the PGA Tour is multifactorial, driving, approach shots, putting, scrambling, and course management all contribute to success.

In addition, a categorical performance label was created using a median split on the Points variable, classifying players into “High” and “Low” performance groups. A K-Nearest Neighbors classifier, Logistic Regression model, and Decision Tree were each trained to predict performance group based on a set of metrics including Fairway Percentage. The best-performing model was Logistic Regression, with an accuracy of 74.2% and an F1-score of 0.73. These results reinforce that driving accuracy is a relevant but not singularly dominant predictor.

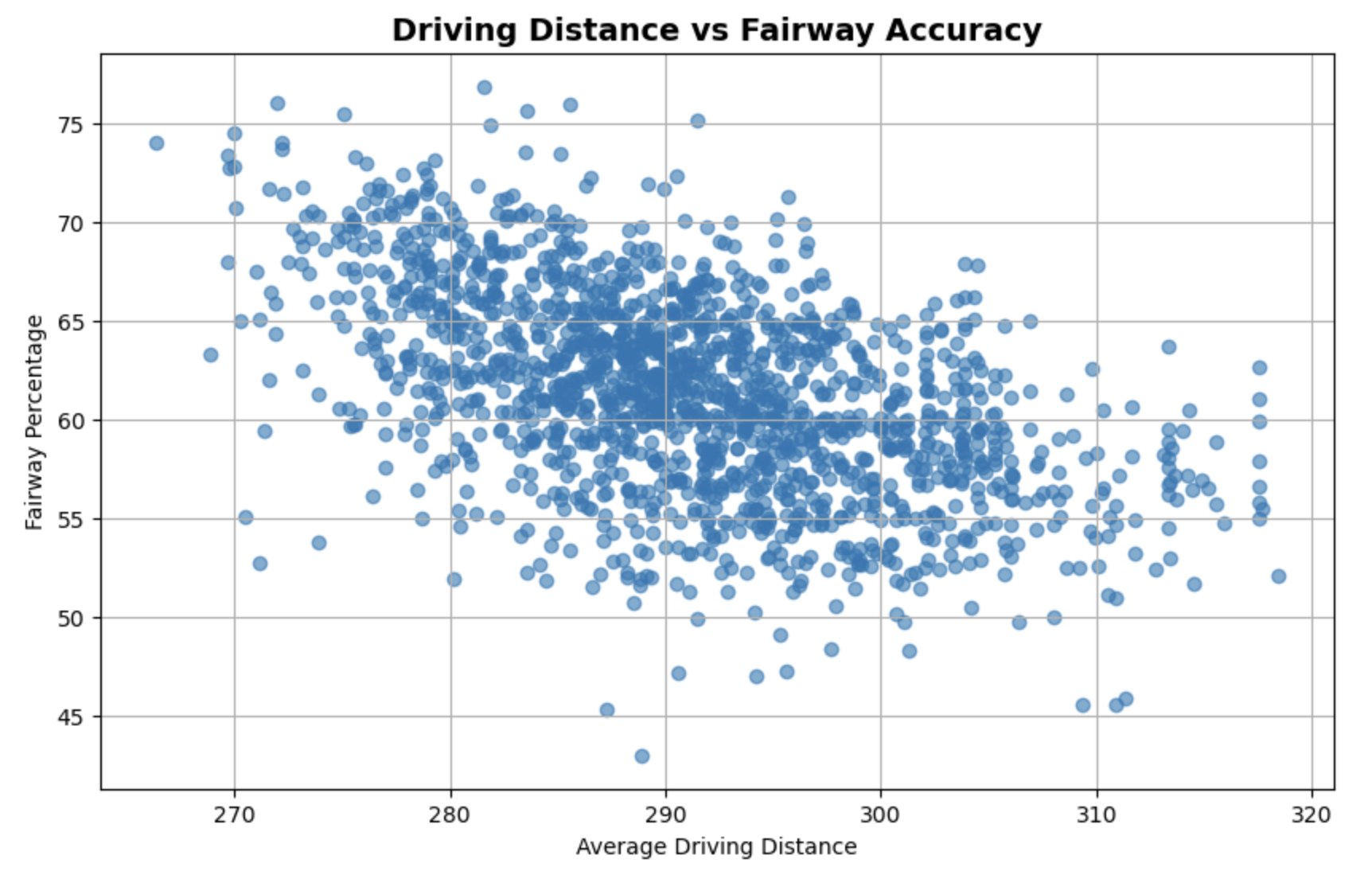
*Interpretation and Broader Context*

From a competitive standpoint, driving accuracy contributes to better approach shot positions and potentially lower scores, which accumulate into more FedEx Cup Points. The statistical evidence presented here supports this connection. However, the relatively low R-squared value highlights that driving accuracy is a piece of the puzzle but not the whole picture. Players with excellent putting and scoring averages can outperform others despite slightly worse tee accuracy.

Still, for aspiring professionals or analysts building performance models, Fairway Percentage is a reliable baseline indicator of a golfer’s consistency and potential for success across the PGA season. Its predictive value may be greater or smaller when combined with other indicators like putting efficiency or scrambling ability.

*3.2 Driving Distance, Accuracy, Performance*

To evaluate whether players with higher driving distance sacrifice accuracy, and how this tradeoff affects performance, we created a scatterplot of average driving distance against fairway percentage using “all\_tour\_data”. The plot, shown in figure 2, shows the relationship between these two key metrics.



The scatterplot reveals a clear negative trend between driving distance and fairway accuracy. The longer a player hits off the tee, the less likely they will be to hit the fairway. This supports our hypothesis of a tradeoff between power and precision.

To determine whether this relationship is statistically significant, we performed a two-sided independent t-test between the fairway accuracy of the longest and shortest drivers. The test produced a t-statistic of –10.78 and a p-value of 2.95e-26 indicating an extremely significant difference in fairway accuracy between the two groups. What we learned from the t-test is that players in the top quartile of driving distance hit significantly fewer fairways on average than those in the bottom quartile, confirming that the distance-accuracy tradeoff is real and robust.

*3.3 Putting Performance and Tournament Success*

Another key area of interest in our study was the role of putting performance in determining overall tournament success, as reflected by FedEx Cup Points. While many aspects of a golfer’s game contribute to performance, putting is frequently cited as a critical skill. This is especially true in high-pressure, scoring-sensitive situations. Our analysis used Average Putts per Round as the primary putting statistic, and Points for cumulative tournament success. We began by visualizing the relationship using a scatterplot with a regression line (Figure 3). This visualization revealed a negative linear trend, suggesting that players who average fewer putts per round tend to earn more FedEx Cup Points across the season. This aligns with the understanding that better putters convert more scoring opportunities and avoid costly three-putts.

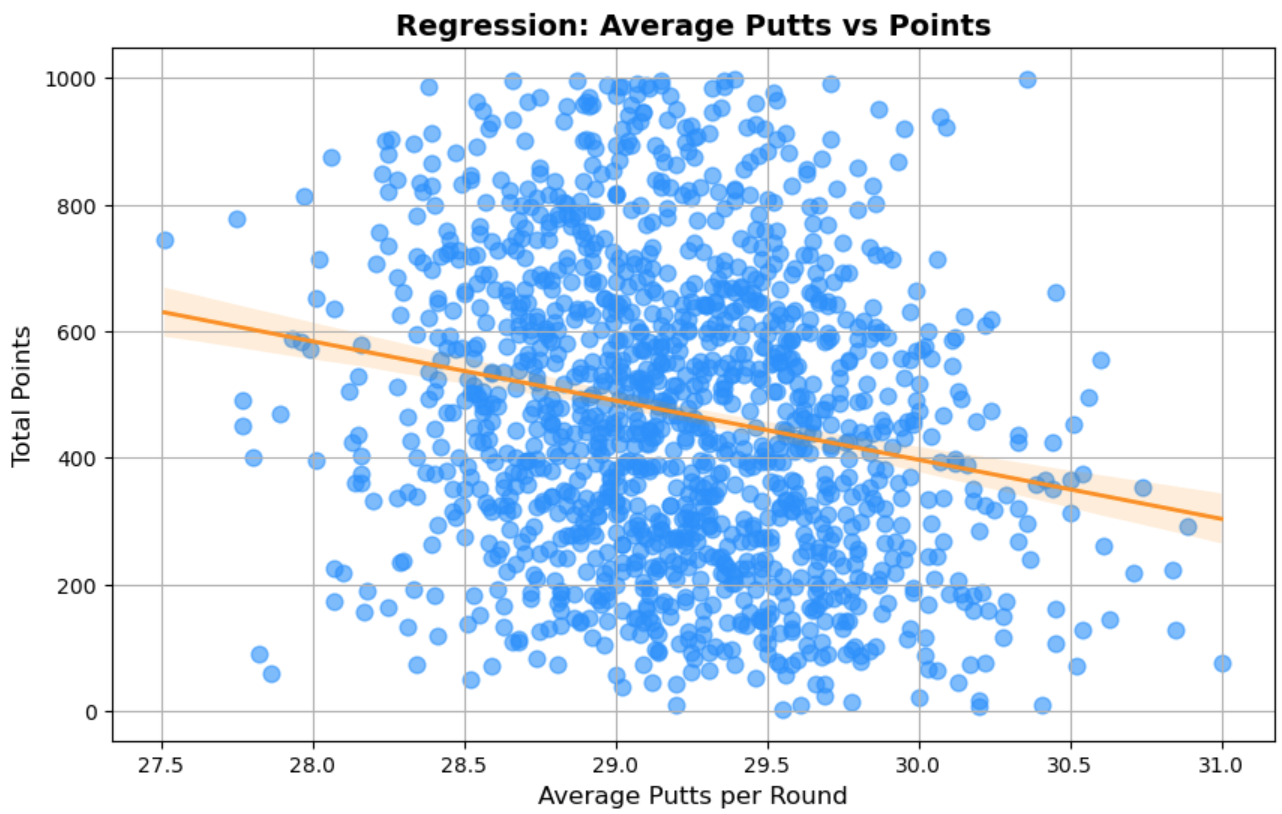


Figure 3: Regression Plot — Average Putts per Round vs. FedEx Cup Points

To statistically validate this relationship, a linear regression was performed with Average Putts as the independent variable and Points as the dependent variable. The regression returned a negative slope coefficient of -51.6, indicating that for each additional putt per round, a player tends to earn approximately 51 fewer FedEx Cup Points over a season. The model’s R-squared was 0.143, meaning roughly 14.3% of the variation in total points can be explained by putting performance alone. These findings reinforce our idea that putting is a differentiating factor at the highest levels of professional golf. While driving and ball striking help set up opportunities, putting ultimately determines how many strokes are converted into scores. Players with fewer putts per round gain a competitive edge, reflected in both leaderboard placements and season-long points accumulation.

To examine how putting contributes to categorical performance outcomes, we split players into “High” and “Low” performers based on median FedEx Cup Points. Using Average Putts along with other features, we trained three classification models—K-Nearest Neighbors, Logistic Regression, and Decision Tree—to predict whether a player would fall into the “High” performance group. Each model showed moderate success in classification, with the Logistic Regression model performing best. Including Average Putts in the feature set improved model performance over models that excluded it. This provides further support that putting is not only correlated with overall success but is also predictive when used alongside other metrics.

*Interpretation and Broader Context*

T he data confirms that putting has a statistically significant and practically meaningful relationship with tournament performance. Players who consistently average fewer putts per round are more likely to be among the top point earners. This makes sense within the structure of modern competitive golf, where courses are designed to test short-game precision and tour events are often won or lost on the greens. Although putting is only one component of overall success, it contributes significantly to scoring efficiency. Its moderate correlation with season-long performance makes it a valuable metric for player evaluation, talent scouting, or strategy development.

*3.4 Average Score Trends Over Time*

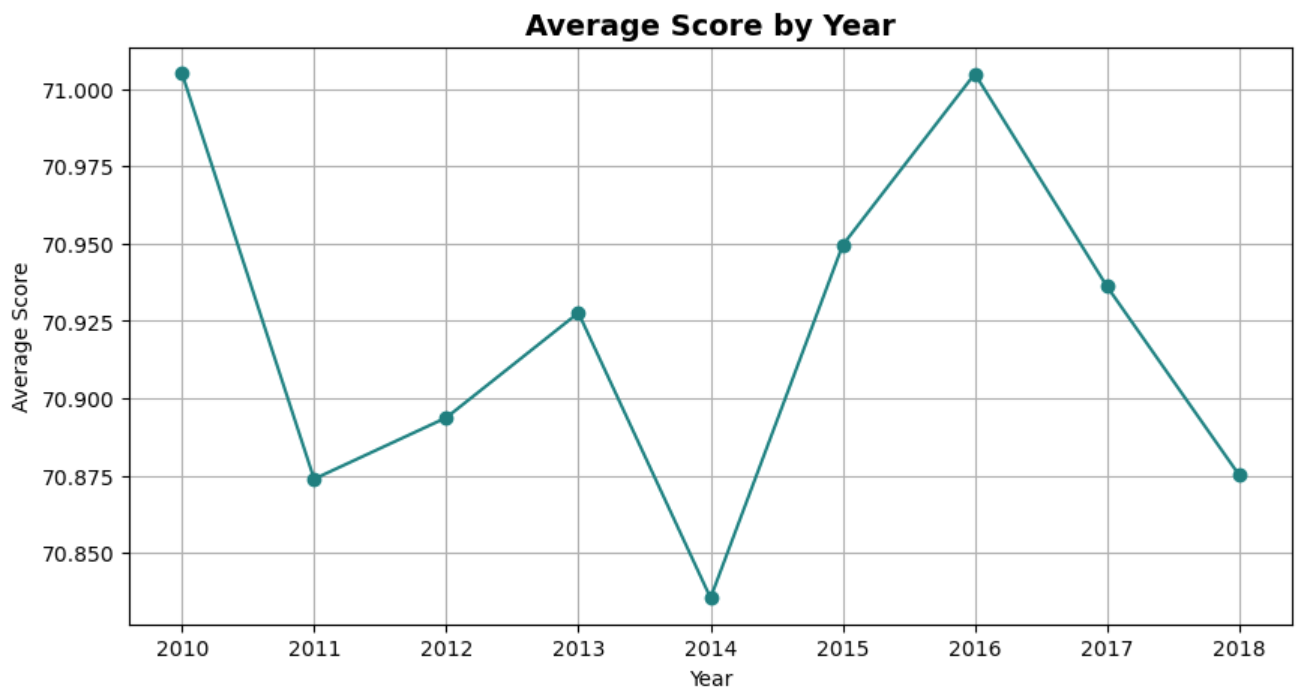
To evaluate performance trends on the PGA Tour, we examined how players’ average scores have changed from 2010 to 2018. Using year-by-year scoring data from the merged dataset, we calculated the mean Average Score for each season and plotted the trend over time (see Figure 4).

Figure 4: Line Plot — Average Score by Year

The results show a gradual decrease in scoring averages over the nine-year span. In 2010, the average score was slightly above 71.3, while by 2018 it had declined to about 70.9. Although the year-to-year variation is small, the trend is consistent and aligns with the broader improvement in player performance due to better training, technology, and data-driven decision-making. A simple linear regression confirmed this trend, with a negative slope of -0.056—suggesting players improved by roughly 0.056 strokes per year. While the R-squared value (~0.27) indicates other variables also influence scoring, this trend demonstrates clear progress in competitive play.

In summary, PGA Tour players have steadily improved scoring efficiency over time, reinforcing the impact of marginal performance gains in professional golf.