Xinyun Yu
2870165
EECS 731
Project 1

Jimmy Wrangler

In this project, given that I am a golfer on KU's golf team, I chose to look at two data sets that are derived from current PGA tour players (*PGA: Professional Golf Association*).

The first data set (*Renamed to PGA_TechData*) has very detailed technical data including "SG_PUTTING_PER_ROUND", "TOTAL_SG:PUTTING", "FAIRWAY_HIT_%" etc, which categorizes the golf game into different techniques a golfer should practise for. The second data set (*Renamed to PGA_GenData*) summarizes how those players perform on the tour on a yearly base for the past 5 years, which is reflected by features "Wins", "Top 10", "Money", "Points", "Average score" etc..

Combining those two data sets could show which technique contributes to the golf game more and thus could help golfers to make wiser choices on the technique they want to improve during practice, especially for those student-athletes who have limited time.

The first step I did was to merge two datasets together using pd.merge() function. The merged dataset (*PGA_DATA*) should only include every detailed technical data including data for putting, driving distance, fairway hit percentage etc. and data that could indicate the players' overall performance including points, wins, top-10s, money etc.. The data that does not belong to either category were deleted from the dataset, so were duplicated ones. Considering the data from 2018 are more recent ones and thus would be more convincing, I chose to only look at data from 2018 by generating a new dataset from the previous one and ignored the data before 2018. PGA_DATA is not overwritten because data before 2018 could be informative for other research (how players have improved their techniques over the years etc.)

Among all of the data that could show how great a player is, the "Average score" from PGA_GenData is the most straightforward one. A player may not have any championship title in the bag yet, but could still be regarded as a great player as long as he has a lower average score when compared to others. Thus, "Average score" will be the feature that I focus on.

I picked out four techniques that are most commonly believed to have direct influence on the golf game, either positively or negatively:

- SG_PUTTING_PER_ROUND: roughly indicate the number of shots you lose due to putting per round. A golfer would prefer a positive number, which means you are benefited from putting. The greater the number is, the better.
- BOGEYS_MADE: Given that you waste one shot for each bogey you make, the less bogeys you have, the better.
- TOTAL_3_PUTTS: A golfer is always allowed to have 2 putts on the green to putt the ball into the hole. Any extra shot would hurt overall score.
- AVG_Driving_DISTANCE: This number indicate how far you can hit with a drive from the tee box. Most golfers believe that, although hitting far could only do good, the driving distance is not the key to the golf game. The rationale behind that is, you could always bounce back by having an approach shot very close to the hole and leave you a chance for a birdie (Birdie: -1shot) even when you had a very bad tee shot.

The four listed data and “Average Score” are plotted with plot() function with “Average score” on the y axis and others on the x axis:

- The graph of “SG_PUTTING_PER_ROUND“” vs. “Average score“” matches what I expected. Although there exist some outliers, the overall trend still indicates that less SG_PUTTING would win you some shots.
- The graph of “BOGEYS_MADE” and “TOTAL_3_PUTTS” differs a little bit than expected. I expected them to have a downward sloping graph while they are pretty evenly distributed. This could be explained by having more birdies could make up to the damages from bogeys and 3-putts. The conclusion here is that bogeys and 3-putts do not hurt as much as we expected.
- The graph “AVG_Driving_DISTANCE” went the opposite way of what I expected. It is clearly shown that the longer driving distance could bring more benefit. This could be explained by leaving shorter distance to approach shot brings more accuracy, which eventually could be transformed to saving a shot on the green. Golfers should be notified here that, hitting gyms more often could help with lowering scores.