GOLF STATS ANALYSIS REPORT

OVERVIEW:

This analysis aimed to explore the relationship between the final golf score and different variables such as fairway percentage, distance, and total putts. The outcome was to produce a model that could predict the final score using statistically essential variables. I analyzed and built two multivariable linear models for two different data sets: the Professional Golf Association (PGA) tour data and my historical golf statistics for every 18 holes. Therefore, after analyzing the differences between the two models, I could learn what variables are not that important to predicting my final score but are essential in the PGA tour data model.

Firstly, I used Pandas, a software package, to understand the datasets and build two multivariable linear models, one for each dataset. During the process, I employed techniques such as exploratory data analysis, data cleaning, and stepwise regression to explore the data types and interesting correlations and find the best model that was the most fitted to the dataset. Then, I compared model results using the values of the two ordinary least squares (OLS) tables produced from the two datasets using pandas. Specifically, I examined the R-squared and F-statistic values as an indicator for model fit and the p-values for statistical significance of variables. This comparison allowed me to understand my golf game better because I saw what I was doing differently compared to male professional golfers; one thing I discovered was that my scrambling percentage was not as statistically crucial in the model as it was for professional players, indicating my scrambling percentage was not consistent while my final score remained relatively consistent.

DATA AND MODEL:

Exploration of data

I used two datasets in my analysis: a PGA tour data set that I found online and my statistics which I named golfStats which I recorded statistics for every 18 holes that I played. The PGA tour data set contained 2,313 professional players across 18 variables. In comparison, the data I collected from my own golf game included 42 rounds across 11 variables. Although there were many variables in the PGA tour data set, some were irrelevant for my analysis. For example, through golf knowledge and common sense, I determined the variables 'PlayerName' and 'Rounds' were not related to predicting the final score.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	PlayerName	Rounds	FairwayPercentage	Year	AvgDistance	gir	AveragePutts	AverageScrambling	AverageScore	Points	Wins	Top10	AveragSGPutts	AveragSGTotal	SG-OTT	SG-APR	SG-ARG	Money
2	Henrik Stenson	60	75.19	2018	291.5	73.51	29.93	60.67	69.617	868		5	-0.207	1.153	0.427	0.96	-0.027	$2,680,487
3	Ryan Armour	109	73.58	2018	283.5	68.22	29.31	60.13	70.758	1,006	1	3	-0.058	0.337	-0.012	0.213	0.194	$2,485,203
4	Chez Reavie	93	72.24	2018	286.5	68.67	29.12	62.27	70.432	1,020		3	0.192	0.674	0.183	0.437	-0.137	$2,700,018
5	Ryan Moore	78	71.94	2018	289.2	68.8	29.17	64.16	70.015	795		5	-0.271	0.941	0.406	0.532	0.273	$1,986,608
6	Brian Stuard	103	71.44	2018	278.9	67.12	29.11	59.23	71.038	421		3	0.164	0.062	-0.227	0.099	0.026	$1,089,763
7	Brian Gay	103	71.37	2018	282.9	64.52	28.25	63.26	70.28	880		6	0.442	0.565	-0.166	0.036	0.253	$2,152,501
8	Kyle Stanley	93	71.29	2018	295.7	71.09	29.89	54.8	70.404	1,198		5	0.037	0.686	0.378	0.298	-0.027	$3,916,001
9	Emiliano Grillo	94	70.16	2018	295.2	68.84	29.04	61.05	70.152	901		5	0.546	1.133	0.364	0.345	-0.122	$2,493,163
10	Russell Henley	77	70.03	2018	293	68.77	29.8	54.33	70.489	569		3	0.167	0.541	0.093	0.467	-0.186	$1,516,438
11	Jim Furyk	50	69.91	2018	280.5	63.19	28.73	62.58	70.342	291		2	0.389	0.412	-0.392	0.179	0.235	$660,010
12	Steve Wheatcroft	60	69.79	2018	288.9	66.57	29.29	61.03	71.631	138		1	-0.128	-0.339	0.112	-0.065	-0.258	$309,656
13	Kevin Streelman	94	69.11	2018	295.1	71.56	29.67	60.93	70.436	673		5	-0.25	0.619	0.439	0.415	0.014	$1,523,642
14	C.T. Pan	104	68.98	2018	292.7	71.2	29.66	56.89	70.457	693		1	-0.067	0.478	0.215	0.267	0.063	$1,881,787
15	David Lingmerth	82	68.93	2018	285.4	63.03	28.5	58.57	71.043	274			0.229	-0.007	0.006	-0.16	-0.081	$616,758
16	Keegan Bradley	98	67.9	2018	299.6	69.18	29.68	56.78	70.303	872		4	-0.358	0.793	0.237	0.888	0.026	$4,069,464
17	Rafa Cabrera Bello	75	67.85	2018	295.1	70.16	29.47	57.98	69.887	784		4	0.273	1.112	0.256	0.487	0.096	$2,449,869
18	Billy Horschel	86	67.8	2018	295.4	71.75	29.46	58.03	70.154	960	1	3	0.392	1.112	0.538	0.352	-0.169	$4,315,200
19	Russell Knox	94	67.7	2018	291.7	69.57	29.7	59.43	70.568	585		3	-0.088	0.383	0.059	0.263	0.149	$1,424,030
20	Ben Crane	65	67.52	2018	281.1	64.88	28.69	63	71.097	267		1	0.332	0.176	-0.302	-0.038	0.184	$620,646
21	Vaughn Taylor	83	67.51	2018	286.1	67.02	29.15	59.91	70.692	445		3	-0.08	0.219	-0.005	0.305	-0.002	$965,691
22	Brian Harman	94	67.14	2018	291.9	67.59	29.29	56.95	70.536	1,056		8	0.273	0.29	0.137	-0.024	-0.096	$2,733,463
23	Sam Ryder	82	66.91	2018	297.3	72.08	29.88	56.47	70.914	442		3	-0.349	0.154	0.203	0.399	-0.099	$1,046,166
24	Ted Potter, Jr.	87	66.83	2018	286	63.03	28.45	57.51	71.024	744	1	1	0.074	-0.094	-0.074	-0.2	0.105	$1,976,198
25	Austin Cook	107	66.76	2018	292.3	66.51	28.72	62.02	70.469	1,060	1	3	0.315	0.569	0.12	-0.045	0.179	$2,448,920
26	Tyler Duncan	97	66.74	2018	294.4	69.65	30.19	52.76	71.04	457		2	-0.566	0.017	0.273	0.476	-0.166	$944,021
27	David Hearn	66	66.63	2018	285.1	68.89	29.58	55.65	71.325	315		2	-0.127	-0.031	-0.17	0.379	-0.113	$622,383
28	Alex Cejka	77	66.49	2018	286.7	63.77	28.52	64	70.675	502		2	0.009	0.312	-0.024	-0.169	0.495	$1,198,541
29	Ian Poulter	73	66.41	2018	293.6	67.01	28.97	57.11	70.593	1,030	1	4	0.223	0.85	0.141	0.435	0.051	$2,714,450

Figure 1 This is the PGA tour data set online from https://www.kaggle.com/jmpark746/pga-tour-data-2010-2018/version/1

The first thing I did to explore the data was to look at the data types of the variables because sometimes the data types were incorrect when loaded into Jupyter Notebook. For example, 'Points' was recognized as an object when it should be an integer.

```
In [9]: pgaTourData.dtypes

Out[9]: Player Name            object
        Rounds                 float64
        Fairway Percentage     float64
        Year                     int64
        Avg Distance           float64
        gir                    float64
        Average Putts          float64
        Average Scrambling     float64
        Average Score          float64
        Points                  object
        Wins                   float64
        Top 10                 float64
        Average SG Putts       float64
        Average SG Total       float64
        SG:OTT                 float64
        SG:APR                 float64
        SG:ARG                 float64
        Money                   object
        dtype: object
```

The next step was to change these incorrect data types to a numerical data type such as integer or float to analyze and build models using these variables. Specifically, the incorrect data types were caused by characters such as commas and dollar marks. Therefore, I removed these characters before I change the datatype.

Figure 2 Exploring the different data types of pgaTourData.

Simple linear regression & Multivariable linear regression

Simple linear regression is used to model the relationship between two continuous variables. Often, the objective is to predict the value of a dependent variable based on the value of an independent variable. A linear regression model would look like y = mx + c where

y = dependent variable

x = explanatory variable

m = slope coefficient for the explanatory variable

c = y-intercept (constant term)

At first, I used a simple linear regression model to visualize the correlation between two variables of pgaTourData.

```
pgaTourData.corr()
```

	Rounds	FairwayPercentage	Year	AvgDistance	gir	AveragePutts	AverageScrambling	AverageScore	Points	Wins
Rounds	1.000000	0.083597	0.078522	0.069599	0.252190	-0.125943	0.227928	-0.313380	0.364636	-0.097741
FairwayPercentage	0.083597	1.000000	-0.111042	-0.533832	0.343115	0.162757	0.236281	-0.170017	0.032302	-0.098652
Year	0.078522	-0.111042	1.000000	0.221102	-0.000044	-0.145524	0.107181	-0.006801	0.037022	0.060289
AvgDistance	0.069599	-0.533832	0.221102	1.000000	0.292965	0.167314	-0.228979	-0.273230	0.281107	0.247126
gir	0.252190	0.343115	-0.000044	0.292965	1.000000	0.549402	0.038068	-0.485030	0.303183	0.267357
AveragePutts	-0.125943	0.162757	-0.145524	0.167314	0.549402	1.000000	-0.549805	0.263041	-0.282488	-0.209932
AverageScrambling	0.227928	0.236281	0.107181	-0.228979	0.038068	-0.549805	1.000000	-0.518057	0.344421	0.248329
AverageScore	-0.313380	-0.170017	-0.006801	-0.273230	-0.485030	0.263041	-0.518057	1.000000	-0.763993	-0.477661
Points	0.364636	0.032302	0.037022	0.281107	0.303183	-0.282488	0.344421	-0.763993	1.000000	0.746627
Wins	-0.097741	-0.098652	0.060289	0.247126	0.267357	-0.209932	0.248329	-0.477661	0.746627	1.000000
Top10	0.211198	0.024990	0.076493	0.226512	0.276390	-0.245181	0.311253	-0.700529	0.823495	0.450695
AveragSGPutts	0.173218	-0.043493	0.023785	-0.136676	-0.231629	-0.720328	0.458220	-0.364817	0.297614	0.216024
AveragSGTotal	0.317011	0.162458	0.029939	0.278923	0.487778	-0.265734	0.517527	-0.962318	0.761594	0.470284
SG-OTT	0.164208	0.179032	0.003330	0.604456	0.582466	0.293925	-0.024871	-0.530624	0.437439	0.244059
SG-APR	0.192764	0.155015	0.009928	0.142217	0.577074	0.133635	0.196124	-0.669370	0.521278	0.348876
SG-ARG	0.112364	0.000356	0.035493	-0.191637	-0.103955	-0.449081	0.613776	-0.390284	0.284049	0.217539
Money	0.278091	0.016237	0.081792	0.329295	0.313770	-0.257832	0.317777	-0.746499	0.955490	0.721482

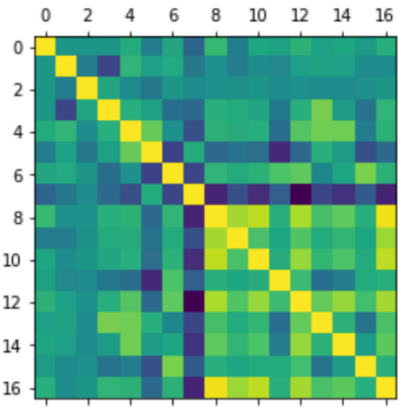

Figure 3 The table showed correlations between all two variables. The coloured correlation table allows one to spot the strength of the correlation easily.

There was a high correlation of -0.962 between the average score and average SG (stoke gained) total. This was an obvious result because these two were both a measure of the score but with a different calculation method. A surprising fact was that fairway percentage did not have a high correlation, around +-0.5 or below, with other variables. As a golfer, fairway percentages are usually considered as one of the factors that could influence the final score.

However, based on my golf knowledge, I knew that there must be a correlation between the variables such as greens in regulation (GIR), fairway percentage and total putts. So, the correlation might not be strong between two variables but between multiple variables. Therefore, I utilized a multivariable regression model as it was more suitable for my goal than a simple linear regression model, which could only use one variable to predict the outcome.

Multivariable linear regression models are used to estimate the relationship between two or more independent variables. They are also used to predict a quantitative outcome. In my analysis, I employed a model to predict the final score.

Formula and Calculation of Multiple Linear Regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p + \epsilon$$

Y = dependent variable

x_i = explanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

Stepwise regression model

Stepwise regression is the step-by-step iterative construction of a regression model that involves selecting independent variables to be used in a final model. It consists in adding or removing potential

explanatory variables in succession based on their importance in the model and testing for statistical significance after each iteration until a best-fitted model is produced.

I used backwards stepwise regression to find my final model, which best fitted my PGA tour dataset based on the F-statistic value. Some variables in the dataset were not statistically significant in predicting the final score. So, I eliminated one or two variables each time and compared the statistics in the OLS Regression Results table.

F-statistic value and p-value

The F-statistic value is the ratio of the mean regression sum of squares divided by the mean error sum of squares. A p-value, or probability value, is a number describing how likely it is that your data would have occurred by random chance that the null hypothesis (there is no correlation between the variables) is true. Therefore, the smaller the p-value, the stronger the evidence that the null hypothesis is not valid.

Specifically, I removed variables with high p-values (>0.05) and compared the entire model through different F-statistic values instead of the R squared values, a value commonly used to determine statistical significance. The R squared value was already one without removing any variables, indicating a perfect fit. However, as I removed variables using my golf knowledge and variables with a high p-value, I found out that the F-statistic value increased, meaning the model fitted the data better. Therefore, I used the F-statistic to decide the model which had the best overall fit. A p-value <0.05 was the threshold that I used to prove that a variable was statistically significant. In my final model, all variables have a p-value of approximately zero. Although it might not be exactly equal to zero, this result indicates the remaining variables are statistically important in predicting the final score.

KEY VARIABLES EXPLANATION:

Greens In Regulation (GIR) – a player hits a green in regulation when their golf ball hits and remains on the putting surface of a hole in as many or fewer than the number of shots prescribed by the par of a hole. (Golf News Net)

Average Scrambling – The percentage of time a player misses the green in regulation but still makes par or better. (Petersson)

SG: OTT (stroke gained on the tee) –This analyses tee shots from a player over the course of a round. The combination of distance and accuracy evaluates how well a player is driving the ball. (TheLines.com)

SG: APR (stroke gained approach shot) – This reflects a player's performance on shots taken from more than 50 yards from the flag, including layup shots. It takes into account the lie players were hitting from, as well as distance and accuracy. ("How Do I Understand Strokes Gained Approach?")

SG: ARG (stoke gained around the green) – Measures how many strokes a player gained or lost on any shot within 30 yards of the green. This statistic does not take into account any shots taken on the green but does factor in shots hit with a putter that are not on the green. (JD)

RESULTS:

OLS Regression Results

Dep. Variable:	AverageScore	R-squared (uncentered):	1.000
Model:	OLS	Adj. R-squared (uncentered):	1.000
Method:	Least Squares	F-statistic:	4.006e+05
Date:	Thu, 26 Aug 2021	Prob (F-statistic):	0.00
Time:	13:48:56	Log-Likelihood:	-187.99
No. Observations:	283	AIC:	406.0
Df Residuals:	268	BIC:	460.7
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Rounds	0.0039	0.003	1.475	0.141	-0.001	0.009
FairwayPercentage	0.0667	0.010	6.657	0.000	0.047	0.086
AvgDistance	0.0680	0.006	11.299	0.000	0.056	0.080
gir	-0.2015	0.023	-8.686	0.000	-0.247	-0.156
AveragePutts	1.8334	0.087	21.174	0.000	1.663	2.004
AverageScrambling	0.1117	0.015	7.672	0.000	0.083	0.140
Points	0.0002	0.000	1.165	0.245	-0.000	0.001
Wins	-0.0231	0.100	-0.232	0.817	-0.219	0.173
Top10	-0.0056	0.027	-0.208	0.836	-0.058	0.047
AveragSGPutts	0.3203	0.818	0.391	0.696	-1.291	1.932
AveragSGTotal	0.0656	0.813	0.081	0.936	-1.535	1.666
SG-OTT	-1.8608	0.820	-2.269	0.024	-3.476	-0.246
SG-APR	-0.6472	0.828	-0.781	0.435	-2.278	0.983
SG-ARG	-0.7371	0.799	-0.923	0.357	-2.310	0.836
Money	-3.877e-08	4.9e-08	-0.791	0.429	-1.35e-07	5.77e-08

Omnibus:	0.342	Durbin-Watson:	1.850
Prob(Omnibus):	0.843	Jarque-Bera (JB):	0.439
Skew:	-0.075	Prob(JB):	0.803
Kurtosis:	2.879	Cond. No.	2.43e+08

OLS Regression Results

Dep. Variable:	AverageScore	R-squared (uncentered):	1.000
Model:	OLS	Adj. R-squared (uncentered):	1.000
Method:	Least Squares	F-statistic:	7.250e+05
Date:	Mon, 16 Aug 2021	Prob (F-statistic):	0.00
Time:	23:35:36	Log-Likelihood:	-196.66
No. Observations:	283	AIC:	409.3
Df Residuals:	275	BIC:	438.5
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
FairwayPercentage	0.0704	0.010	7.102	0.000	0.051	0.090
AvgDistance	0.0728	0.005	13.271	0.000	0.062	0.084
gir	-0.1676	0.021	-8.090	0.000	-0.208	-0.127
AveragePutts	1.6667	0.066	25.407	0.000	1.538	1.796
AverageScrambling	0.1376	0.013	10.692	0.000	0.112	0.163
SG-OTT	-1.9845	0.123	-16.171	0.000	-2.226	-1.743
SG-APR	-0.6790	0.105	-6.437	0.000	-0.887	-0.471
SG-ARG	-0.8721	0.188	-4.637	0.000	-1.242	-0.502

Omnibus:	1.032	Durbin-Watson:	1.793
Prob(Omnibus):	0.597	Jarque-Bera (JB):	0.874
Skew:	-0.133	Prob(JB):	0.646
Kurtosis:	3.056	Cond. No.	2.04e+03

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 2.04e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 4 Comparison of the OLS Regression Results tables of the pgaTourData model before and after employing stepwise regression model

Final model of pgaTourData:

Final Score = 0.0704FairwayPercentage + 0.0728AvgDistance - 0.1676gir + 1.6667AveragePutts + 0.1376AverageScrambling – 1.9845SG-OTT - 0.6790SG-APR – 0.8721SG-APR

The left-side table, the OLS Regression Results table before using the stepwise regression, has an F-statistic value of 4.006e+05. Whereas the right-side table, the results of the final model, has an F-statistic value of 7.250e+05. The larger the F-statistic value, the more fit the model is to the data points. Therefore, this indicates the elimination of variables improved the model. Also, from the original table, some variables have high p-values, such as 'AveragSGTotal' with a p-value of 0.936. As mentioned previously, if the p-value is higher than the 0.05 threshold, the variable is not statistically significant. Therefore, these variables were eliminated as the first few options. In comparison, in the final table of the best fit model, the p-values of the remaining variables are all approximately 0. This shows how they are all extremely close to zero, meaning they are all statistically significant to the model in predicting the final score.

OLS Regression Results

Dep. Variable:	FinalScore	R-squared (uncentered):	0.999
Model:	OLS	Adj. R-squared (uncentered):	0.999
Method:	Least Squares	F-statistic:	6948.
Date:	Thu, 26 Aug 2021	Prob (F-statistic):	6.04e-51
Time:	10:29:28	Log-Likelihood:	-83.060
No. Observations:	41	AIC:	183.7
Df Residuals:	33	BIC:	197.4
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
GIR	-0.0945	0.044	-2.140	0.040	-0.184	-0.005
Fairway	0.0272	0.024	1.152	0.258	-0.021	0.075
Putts	1.0647	0.162	6.552	0.000	0.734	1.395
Distance	0.1903	0.022	8.641	0.000	0.145	0.235
Scrambling	0.0132	0.018	0.723	0.475	-0.024	0.050
SG:OTT	-1.8245	0.265	-6.879	0.000	-2.364	-1.285
SG:APR	-0.5619	0.252	-2.228	0.033	-1.075	-0.049
SG:ARG	-0.2675	0.259	-1.034	0.309	-0.794	0.259

Omnibus:	2.319	Durbin-Watson:	1.923
Prob(Omnibus):	0.314	Jarque-Bera (JB):	1.384
Skew:	0.412	Prob(JB):	0.500
Kurtosis:	3.363	Cond. No.	252.

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Final Model of golfStats:

Final Score = -0.0945GIR + 0.0272Fairway + 1.0647Putts + 0.1903Distance + 0.0132Scrambling – 1.8245SG:OTT – 0.5619SG:APR – 0.2675SG:ARG

The table on the left is the OLS Regression Results of my personal golf stats dataset which I named golfStats. I included the remaining variables

Figure 5 The OLS Regression Results table of golfStats

from the final pgaTourData model after using stepwise regression to make some comparisons. The R-squared and adjusted R-squared values are both 0.999, indicating the model fits the data extremely well. However, the F-statistic value of the golfStats model is a lot smaller than the one of the pgaTourData model. Also, it is interesting how the variable 'scrambling' has a p-value of 0.475 which is relatively high compared to the p-value of the variable 'AverageScrambling", 0.000, of pgaTourData model. This presents how according to my personal data, scrambling percentage is not statistically significant in predicting the final score. Another interesting result is how the p-value of the variable 'distance' remains to be 0.000. This stresses the importance of distance in shooting a good score in golf, which is something my coach often emphasizes but I tend to ignore.

CONCLUSION:

Overall, I have achieved my goal of creating a model using stepwise regression to predict a round's final score. At the same time, it is important that the limitations of both results are taken into consideration. Both datasets are biased towards certain groups: the first dataset, pgaTourData, is biased towards male professional players; the other dataset, golfstats, is biased towards me. Also, pgaTourData has a lot more data points than golfstats. Therefore, the comparison between the two datasets cannot be made directly. However, accounting for this bias while making the comparisons allowed me to learn more golf knowledge and better understand my golf game.

Future analysis could be done in reversing the prediction. This means that given the final score, the model can produce an exact value for each variable based on the dataset that I used. This will make an impact on improving one's golf game as one is able to have a goal for each aspect before going into a round. Also, it would be interesting to build another model based on LPGA (ladies professional golf association) data and compare it with PGA data.

KEY VARIABLES EXPLANATION BIBLIOGRAPHY:

Golf News Net. "What Is a Green in Regulation in Golf? What Does That Term Mean?" *Golf News Net: What You Need to Know about Golf*, 15 July 2021, thegolfnewsnet.com/golfnewsnetteam/2021/07/15/what-is-a-green-in-regulation-in-golf-what-does-that-term-mean-123465/\.

"How Do I Understand Strokes Gained Approach?" *Arccos Golf*, support.arccosgolf.com/hc/en-us/articles/360037682932-How-do-I-understand-Strokes-Gained-Approach-. Accessed 29 Aug. 2021.

JD. "ON: STROKES GAINED EXPLAINED - JD." *Medium*, 9 Apr. 2019, medium.com/@jamesmazzolajd/on-strokes-gained-explained-1e92758ef93d.

Petersson, Thomas. "Scrambling in Golf – How to Use It to Improve." *Anova.Golf*, 15 Jan. 2020, anova.golf/scrambling-in-golf.

TheLines.com. "What Is Strokes Gained In Golf? | PGA Tour Betting Strategy." *The Lines*, 16 Nov. 2020, www.thelines.com/betting/golf-betting/what-is-strokes-gained.

Code:

https://github.com/Aliendoo/Golf-Stats.git