Statistical Methods

PGA Golf Tournament Analysis

Submitted by Pravallika Kalidindi

(M12823119)

Objective

This report aims to answer some of the questions raised by Chriss Higgins and Colin Mcdougall regarding Golf using The Professional Golfers Association (PGA) Golf data.

The methods used to answer questions include descriptive statistics, paired t-test, two sample t-test etc.

The questions answered are as follows-

* Are scores different from the first day to the last day?
* Are young people doing better than those who are older?
* Do long hitters have low scores

Introduction

Colin McDougall, a golf enthusiast and former NCAA golfer at the University of Tampa, was watching the final round of a recent golf tournament with Chris Higgins, the resident statistician at the Richard Ivey School of Business. Higgins, an avid golfer himself, was amazed at the low scores being made on this particular Sunday. McDougall, however, was not impressed. He explained that scores on Sundays are always lower than scores on Thursdays for the benefit of television ratings. “Fans want to see golfers making birdies so the golf courses are set up to be easier on Sunday,” said McDougall.

Higgins countered that all the pressure facing the golfers on Sunday would surely raise their scores, and the television coverage would increase the pressure. McDougall was not convinced. “Too bad we don’t have the data” he said. It would make an interesting study.

In this particular tournament, a young unknown was leading the field. McDougall believed that the younger players were taking over. McDougall, who was over 40 himself, lamented, “Once you hit 40 you are simply not competitive anymore.”

As the final round progressed, McDougall and Higgins discussed questions that included the following:

1. Are scores different from the first day to the last day?
2. Are scores different across the four rounds?
3. Are young people doing better than those who are older?
4. Do long hitters have lower scores?
5. How important is driving accuracy in determining one’s score?
6. Do people putt for “dough” and drive for show?

This report will answer some of these questions using techniques like descriptive statistics, , paired t-test, two sample t-test etc.

Data

The data for analysis has been obtained by the Professional Golfers Association (PGA). The data contained the scoring statistics of players. data were modified to show 270 unique golfers and a summary of what they accomplished in all tournaments where they played four rounds (i.e., made the cut). The data had the following columns

|  |  |
| --- | --- |
| Column Name | Description |
| PlayerNumber | Unique number identifying the player |
| Age | Player age |
| FedExCupPoints | FedEx Cup points |
| Money | Average money won per tournament |
| Round1Score | Average strokes in Round 1 |
| Round2Score | Average strokes in Round 2 |
| Round3Score | Average strokes in Round 3 |
| Round4Score | Average strokes in Round 4 |
| Total Strokes | Average strokes per tournament |
| Average Drive | Average driving distance |
| Drive\_Rank | Driving rank |
| Percent\_Birdie\_when\_GIR | Percent birdies made when green is hit in regulation |
| Percent\_Fairways | Percent of drives in fairway |
| Percent\_GIR | Percent greens hit in regulation |
| Putt\_Round | Average Putts per round |
| Percent\_10foot | Percent of putts inside 10’ made |
| Percent\_Outside10 | Percent of putts outside 10’ made |

Background

Virtually every Professional Golfers Association (PGA) tournament is four rounds of golf with the golfer having the lowest total score being declared the winner. Most tournaments start with approximately 156 players. After two rounds of golf, the number of golfers is reduced (usually by 50 per cent) based on the total scores over those two rounds. Those players who qualify to play the final two rounds (on Saturday and Sunday) are said to have “made the cut.”

Golf courses have three types of holes: par 3s, par 4s and par 5s. Making par on the hole means the golfer took the expected number of shots. For example, on a par 3 hole, a golfer is expected to score 3, which includes the first shot to the green and two putts. On a par 4, the golfer is expected to take two shots to hit the green and then two putts. On a par 5, the golfer is expected to take three shots to hit the green and then two putts. A golfer is said to have hit the green in regulation (GIR) when the golfer gets on the green in one shot on a par 3 hole, two shots on a par 4 hole and three shots on a par 5 hole.

On par 4 and par 5 holes, the first shot off the tee is called a drive. Golfers try to land their ball in the fairway, where the grass is cut short and it is easier to hit the next shot. The area off the fairways, where the grass is longer, is called the rough. A high percentage of drives in the fairway (percent fairways) is considered good and typically leads to lower scores. On par 3 holes, the first shot is not considered a drive.

Putting is hugely important in golf. Putting takes place on the greens. Two putts are considered the norm, so sinking the first putt obviously leads to lower scores. Given there are 18 holes in a round of golf, the expected number of putts is 36. Any fewer putts is good.

Analysis

1. **Question 1** - Are scores different from the first day to the last day?

To answer this question we will use a paired t-test between Round 1 and Round 4 scores. A paired t-test is used to compare two population means where you have two samples in which observations in one sample can be paired with observations in the other sample. In this context since we can pair Round 1 and Round 4 scores as scores obtained by the same player in two different rounds, we can use a paired t-test.

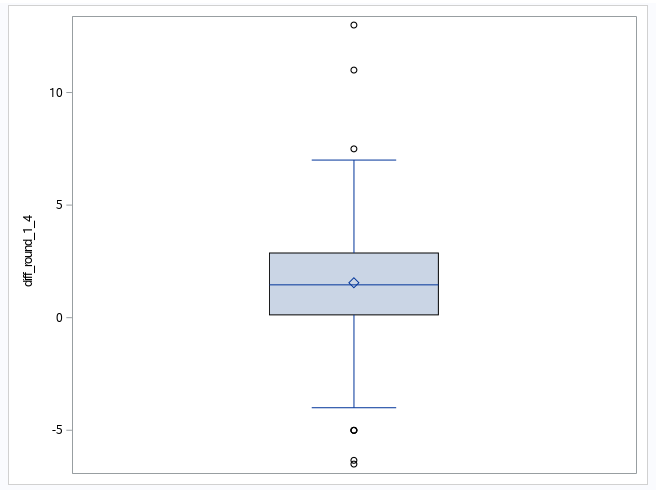
**Method** – paired t-test

**Assumptions** **of a paired t-test** –

* Assumption 1 : There should be no significant outliers in the differences between the two related groups.
* Assumption 1 : The distribution of the differences in the dependent variable between the two related groups should be approximately normally distributed

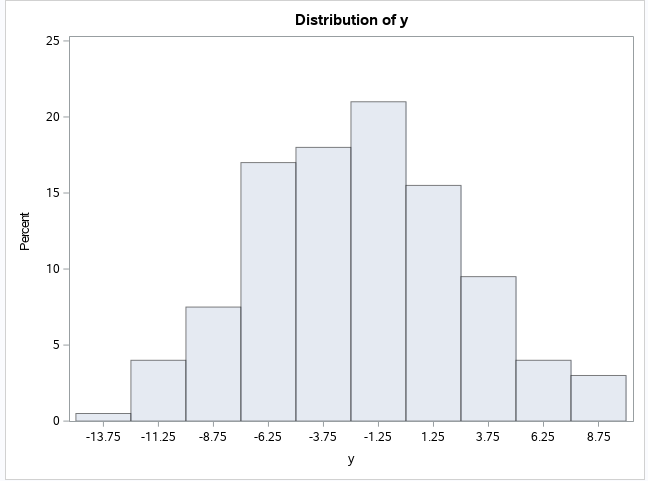
**Test of Assumptions**

* Test for Assumption 1 – In order to test assumption 3 , we will draw a box plot for the difference in round 1 and round 4 scores.



From the box plot above, we can observe a few outliers. But since the number of outliers is not a lot compared to the numbers of observations we have (270), we can still use a paired t-test.

* Test for assumption 2– To test assumption 4, we will draw a histogram for difference in round 1 and round 4 scores and see if it follows an approximate normal distribution.



From the histogram above distribution of difference in round 1 and round 4 scores are approximately normal.

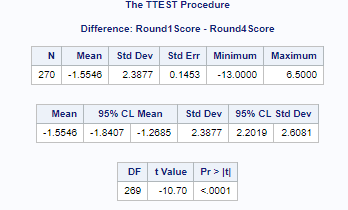
**Analysis-**

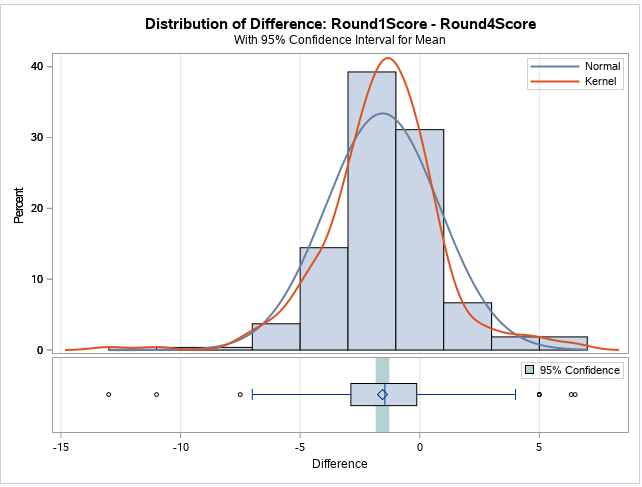
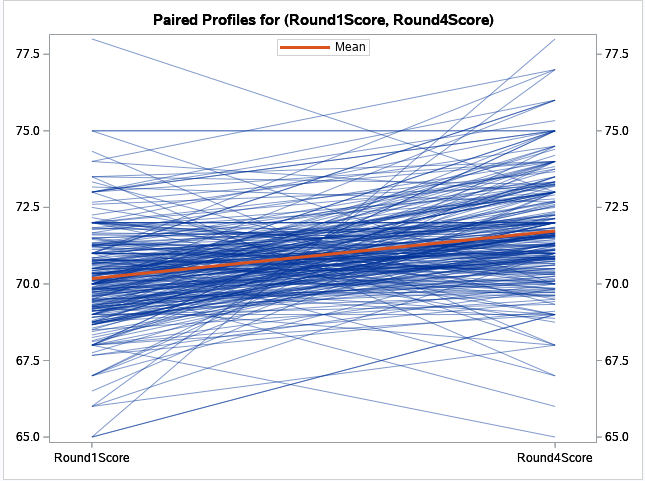
Null Hypothesis (H0) = Mean of round 1 score = Mean of round 4 score

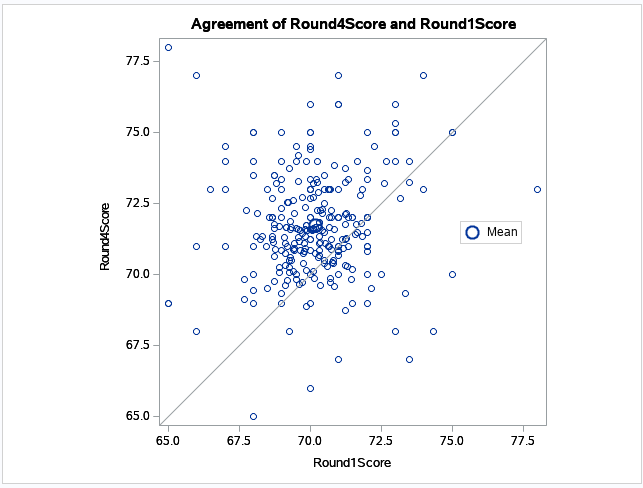
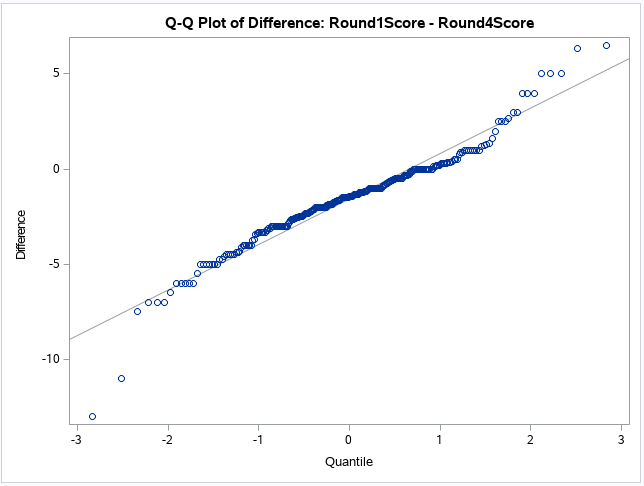
Alternate Hypothesis = Mean of round 1 score is not equal to Mean of round 4 score.

Confidence level = 0.05

After performing a paired t-test for Round 1 and Round 4 scores, the following results are obtained.



**Interpretation**

From the results presented above, p-value <0.0001 . Since p-value<alpha , we can reject the null hypothesis that there is no difference in scores on day 1 and day 4 .

Scores obtained on day 1 and day 4 are significantly different.

1. **Question 2 -** Are young people doing better than those who are older?

To answer this question we will consider players aged less than 40 as Young players and players with age 40 and above as old players. We are also using TotlaStrokes column in our data as an estimate for players overall score.

**Method –** T-Test

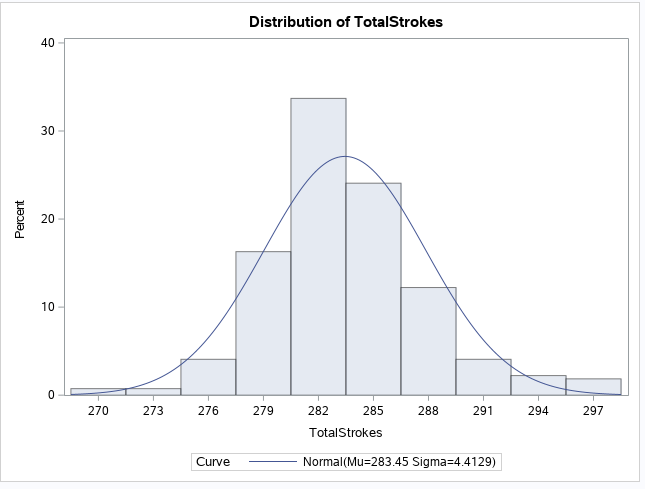
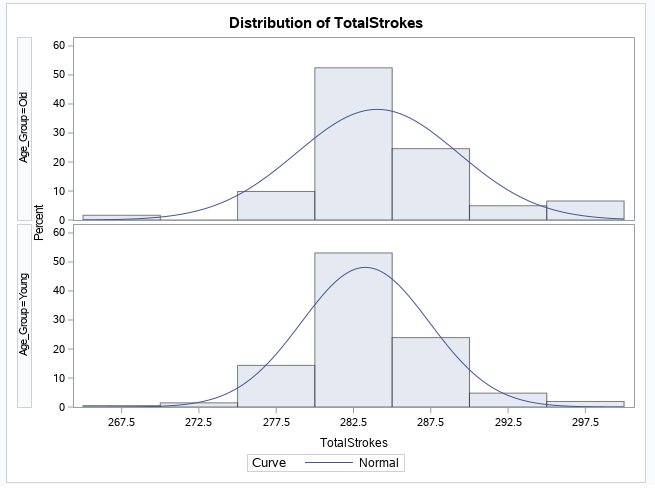
We will use a normal t-test and not a paired t-test because we cannot pair observations for the age group young and old.

**Assumptions of T-Test**

* Assumptions 1 - Dependent variable(Total Strokes) has a normal distribution.
* Assumption2 - Dependent variable has the same variance, σ2, in each group (as though the distribution for group A were merely shifted over to become the distribution for group B, without changing shape).

**Test of Assumptions –**

To test the above assumptions we will use histograms and see if the distributions follow normal and if they have similar variance.

From the graphs above, we can assume that our dependent variable follows an approximate normal distribution with similar variance for both age groups.

**Analysis-**

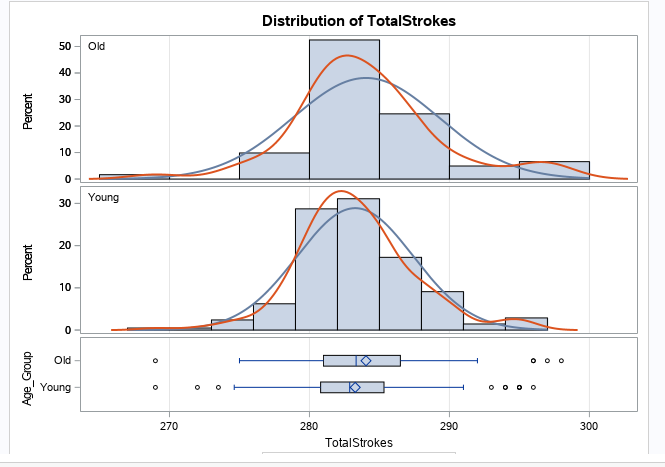
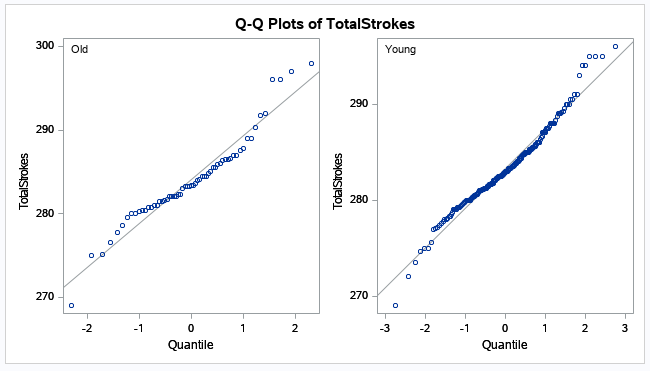
Null Hypothesis (H0) = Mean score for young people <= Mean score for old people

Alternate Hypothesis = Mean score for young people > Mean score for old people.

Confidence level = 0.05

After performing a t-test for TotalStrokes, the following results are obtained.



**Interpretation**

**For equality of variances-**

Null Hypothesis : Variance of scores of old players = variance of scores of young players .

Alternate Hypothesis : Variance of scores of old players != variance of scores of young players

Confidence level =0.05

From the results presented above, p-value for equality of variances is 0.0177 . Since p-value<alpha , we can reject the null hypothesis that Variance of scores of old players = variance of scores of young players. Therefore the two groups have unequal variance.

Taking into consideration, the p value for unequal variance is 0.1460. Since p-value>alpha , we fail to reject the null hypothesis that Mean score for young people <= Mean score for old people.

Therefore old people have higher scores than young people.

Since a player wins if he has the least score, we can conclude that younger players are doing better than older players.

1. **Question 3 -** Do long hitters have lower scores?

To answer this question we are using TotlaStrokes column in our data as an estimate for players overall score. We will use column Average Drive to classify players as long hitters and short hitters. Players with drive distance greater than mean drive distance are classified as long hitters and players with drive distance less than mean drive distance are classified as short hitters.

**Method –** T-Test

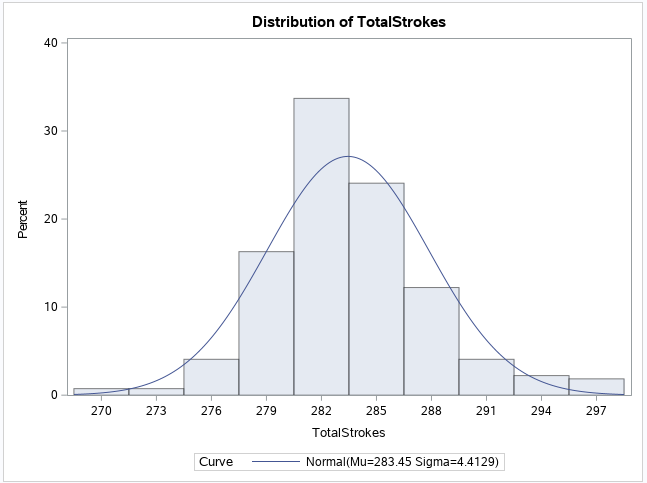
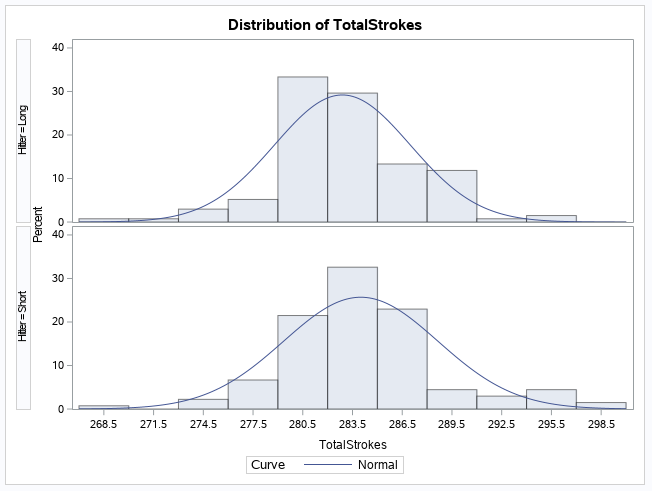
We will use a normal t-test and not a paired t-test because we cannot pair observations for the long hitters and short hitters.

**Assumptions of T-Test**

* Assumptions 1 - Dependent variable(Total Strokes) has a normal distribution.
* Assumption2 - Dependent variable has the same variance, σ2, in each group (as though the distribution for group A were merely shifted over to become the distribution for group B, without changing shape).

**Test of Assumptions –**

To test the above assumptions we will use histograms and see if the distributions follow normal and if they have similar variance.

From the graphs above, we can assume that our dependent variable follows an approximate normal distribution with similar variance for both long hitters and short hitters.

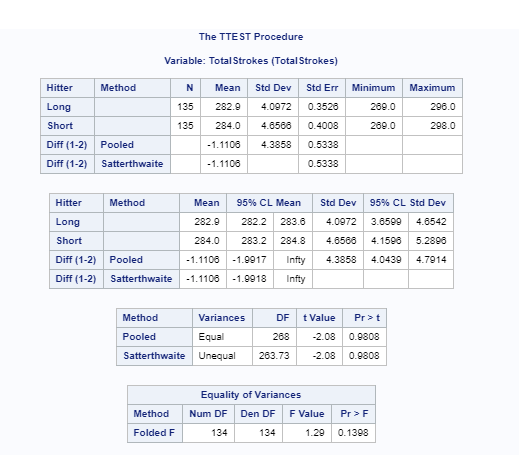
**Analysis-**

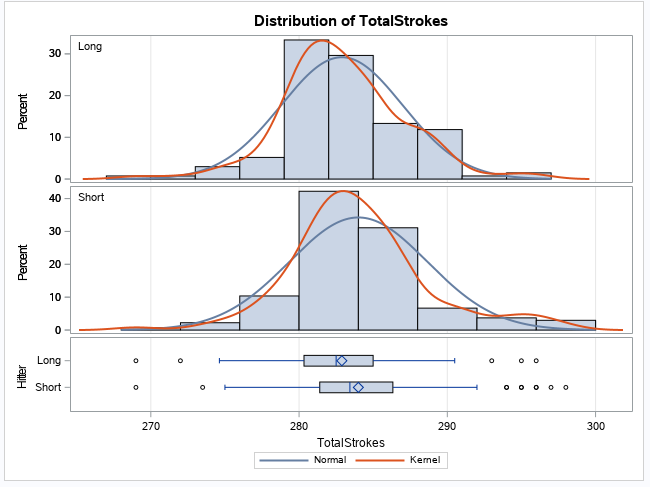
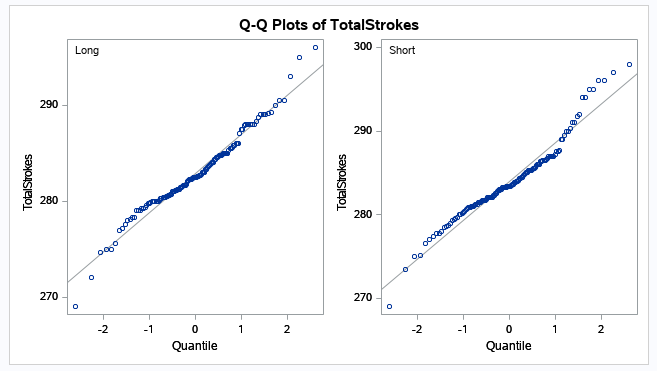
Null Hypothesis (H0) = Mean score for long hitters <= Mean score for short hitters

Alternate Hypothesis = Mean score for long hitters > Mean score for short hitters.

Confidence level = 0.05

After performing a t-test for TotalStrokes, the following results are obtained.



**Interpretation**

**For equality of variances-**

Null Hypothesis : Variance of scores of long hitters = variance of scores of short hitters .

Alternate Hypothesis : Variance of scores of long hitters != variance of scores of short hitters

Confidence level =0.05

From the results presented above, p-value for equality of variances is 0.1398 . Since p-value>alpha , we fail to reject the null hypothesis that Variance of scores of long hitters = variance of scores of short hitters . Therefore the two groups have equal variance.

Taking into consideration, the p value for equal variance is 0.9808. Since p-value>alpha , we fail to reject the null hypothesis that Mean score for long hitters <= Mean score for short hitters.

Therefore long hitters do have lower scores than short hitters.

**Appendix**

SAS code for all above questions can be found below. The output graphs and tables are provided above along with the relevant questions.

* **Data import-**

**PROC** **IMPORT** OUT= WORK.PGA DATAFILE= "/folders/myfolders/GolfingData.xlsx"

DBMS=XLSX REPLACE;

SHEET="Sheet1";

GETNAMES=YES;

**RUN**;

* **Question 1 –**

1. Creating new data set with new column = Round 4 Score - Round 1 Score

**DATA** PGA\_NEW;

SET WORK.PGA;

DIFF\_ROUND\_1\_4= ROUND4SCORE-ROUND1SCORE;

**RUN**;

1. Creating box plot and histograms for new column DIFF\_ROUND\_1\_4 to test for assumptions.

**PROC** **SGPLOT** DATA=PGA\_NEW;

VBOX DIFF\_ROUND\_1\_4;

**RUN**;

**PROC** **UNIVARIATE** DATA=PGA\_NEW ;

VAR DIFF\_ROUND\_1\_4 ;

HISTOGRAM;

**RUN**;

1. Paired t-test for Round 1 score and Round 4 score

**PROC** **TTEST** DATA=WORK.PGA;

PAIRED ROUND1SCORE\*ROUND4SCORE;

**RUN**;

* **Question 2 –**

1. Create a new column Age\_Group for partitioning Young players and Old players.

**DATA** AGE\_SCORE;

SET WORK.PGA;

IF AGE<**40** THEN AGE\_GROUP="YOUNG";

ELSE AGE\_GROUP="OLD";

KEEP AGE\_GROUP TOTALSTROKES;

**RUN**;

1. Creating histograms for column TotalStrokes and TotalStrokes by Age\_Group to test for assumptions.

**PROC** **UNIVARIATE** DATA=AGE\_SCORE ;

CLASS AGE\_GROUP;

VAR TOTALSTROKES ;

HISTOGRAM/NORMAL;

**RUN**;

**PROC** **UNIVARIATE** DATA=AGE\_SCORE ;

VAR TOTALSTROKES ;

HISTOGRAM/NORMAL;

**RUN**;

1. T-test for TotalStrokes for young players and old players.

**PROC** **TTEST** SIDES=U DATA=AGE\_SCORE;

CLASS AGE\_GROUP;

VAR TOTALSTROKES;

**RUN**;

* Question 3

1. Creating new column Hitter to classify players as Long hitters and Short hitters.

**PROC** **SUMMARY** DATA=WORK.PGA NWAY;

VAR AVERAGEDRIVE: ;

OUTPUT OUT=MEANDRIVE (DROP=\_:) MEAN=;

**RUN**;

**DATA** PGA\_LONG\_SHORT\_HITTERS;

SET WORK.PGA;

LENGTH HITTER $5.;

MEANAVERAGEDRIVE=**279.935**;

IF AVERAGEDRIVE>=MEANAVERAGEDRIVE THEN HITTER="LONG";

ELSE HITTER="SHORT";

**RUN**;

1. Creating histograms for column TotalStrokes and TotalStrokes by Hitter group to test for assumptions.

**PROC** **UNIVARIATE** DATA=PGA\_LONG\_SHORT\_HITTERS ;

CLASS HITTER;

VAR TOTALSTROKES ;

HISTOGRAM/NORMAL;

**RUN**;

**PROC** **UNIVARIATE** DATA=PGA\_LONG\_SHORT\_HITTERS ;

VAR TOTALSTROKES ;

HISTOGRAM/NORMAL;

**RUN**;

1. T-test for TotalStrokes for long hitters and short hitters

**PROC** **TTEST** SIDES=U DATA=PGA\_LONG\_SHORT\_HITTERS;

CLASS HITTER;

VAR TOTALSTROKES;

**RUN**;