Assessing Golfer Performance on the PGA Tour

Luke Chamberlain

x13521763

National College of Ireland

BSc (Hons) in Technology Management (*Data Analytics*)

2016 / 2017

10/05/2017

Declaration Cover Sheet for Project Submission

SECTION 1 *Student to complete*

Name: Luke Chamberlain
Student ID: X13521763
Supervisor: Dr Eugene F.M. O'Loughlin

SECTION 2 Confirmation of Authorship

The acceptance of your work is subject to your signature on the following declaration:

I confirm that I have read the College statement on plagiarism (summarised overleaf and printed in full in the Student Handbook) and that the work I have submitted for assessment is entirely my own work.

Signature: Luke Chamberlain

Date: 09/05/2017

NB. If it is suspected that your assignment contains the work of others falsely represented as your own, it will be referred to the College's Disciplinary Committee. Should the Committee be satisfied that plagiarism has occurred this is likely to lead to your failing the module and possibly to your being suspended or expelled from college.

Complete the sections above and attach it to the front of one of the copies of your assignment,

What constitutes plagiarism or cheating?

The following is extracted from the college's formal statement on plagiarism as quoted in the Student Handbooks. References to "assignments" should be taken to include any piece of work submitted for assessment.

Paraphrasing refers to taking the ideas, words or work of another, putting it into your own words and crediting the source. This is acceptable academic practice provided you ensure that credit is given to the author. Plagiarism refers to copying the ideas and work of another and misrepresenting it as your own. This is completely unacceptable and is prohibited in all academic institutions. It is a serious offence and may result in a fail grade and/or disciplinary action. All sources that you use in your writing must be acknowledged and included in the reference or bibliography section. If a particular piece of writing proves difficult to paraphrase, or you want to include it in its original form, it must be enclosed in quotation marks and credit given to the author.

When referring to the work of another author within the text of your project you must give the author's surname and the date the work was published. Full details for each source must then be given in the bibliography at the end of the project

Penalties for Plagiarism

If it is suspected that your assignment contains the work of others falsely represented as your own, it will be referred to the college's Disciplinary Committee. Where the Disciplinary Committee makes a finding that there has been plagiarism, the Disciplinary Committee may recommend

- that a student's marks shall be reduced
- that the student be deemed not to have passed the assignment
- that other forms of assessment undertaken in that academic year by the same student be declared void
- that other examinations sat by the same student at the same sitting be declared void

Further penalties are also possible including:

- suspending a student college for a specified time,
- expelling a student from college,
- prohibiting a student from sitting any examination or assessment.,
- the imposition of a fine,
- the requirement that a student to attend additional or other lectures or courses or undertake additional academic work.

Abstract

For many Professional Golfers success can be rare on the PGA Tour. Attaining membership on the PGA Tour is the pinnacle of most Professional Golfer's careers. Many Professionals pursue and fail in the search for success on the PGA Tour at the expense of leaving their families and constant travelling. At the end of the day Professional Golf is a full time job, and while many earn millions every year, some players go week after week with no earnings from missing the Cut.

This dissertation will attempt to measure and predict what it takes to be successful on the PGA Tour by analysing data from Shotlink. The analysis and predictions will be completed through predictive analysis methods and various machine learning models. The expected results are intended to provide professional golfers with the key components of what it takes to be successful.

Additionally, the dissertation will also aim to answer several critical questions frequently asked among the world of golf.

Acknowledgements

I would like to sincerely thank my Supervisor Dr Eugene F.M. O'Loughlin for his consistent advice and guidance throughout competition of this Project.

I would also like to express gratitude to ShotLink as without them granting me access to their exceptional data this project would not be possible.

Special thanks also go to family and friends for their thorough proof reading of this dissertation and for keeping me sane throughout this project.

Glossary of Terms

ShotLink / Database – Database which all data is been gathered from

Project – Research Dissertation to be completed

Intended Audience – Professional Golfers and Golf Researchers

ShotLink Management / Data Stewards – Team in charge of dealing with Data Requests, Administering Data and approving final dissertation

Exported Data – Data exported from "ShotLink" used for research purposes

PGA Tour – Top Tier Golf Tour Which Professional Golfers are a member of in United States of America

User – Person who has access to data

GUI – Graphical User Interface

SPSS – A Statistical Analysis software package owned by IBM

RStudio – Platform used for R

R – Statistical Programming Language

Data Analysis Toolpak – Microsoft Excel program used for analysis

GIR – Green in Regulation

Table of Contents

Table of Figures

1. Introduction

Golf is a game of a hitting a little white ball into a hole, sounds easy right? Think again. Golf is extremely complex with many factors making it one of the most technical sports in the world. Golf has always been associated with the upper class and the older generation. In recent times it has become a young person's sport with high levels of athleticism. Membership on the PGA Tour is what professional golfers aspire to at the outset. Success on the PGA Tour gives players the potential to win Millions. While this all sounds extravagant, there is also another side to the beautiful game of golf. Some golfers spend years trying to make it to the top, spending all their money and time on getting there and end up failing the majority of the time. There is such a small percentage of success for players at a professional level. It is said that roughly 80% of all prize money in professional golf is distributed among the top 300 golfers the world. This is incredible considering there are approximately 30,000 professional golfers in the PGA world rankings. It is also said that there is somewhere near 60 million people in the world who play golf. If you crunch the numbers, it shows that there is a very small chance of making it to the top.

1.1 Background

To many, the game of golf is as clear as rocket science. While they may be able to grasp the concept that a round is played across the span of 18 holes, that just scratches the surface. As an avid golfer, common questions I come across are such as; "what is a birdie or par?", "why do you need so many clubs?" and of course "why is there so many rules?" This will all be explained in detail later on.

This shows the game is not as simple as just hitting the ball into the hole. There are much more complicated fundamentals and technicalities that make a good golfer such as; the swing, the mind and composure, consistency, connections and luck. While these characteristics may not be measureable I will attempt to project what it does take to become a successful golfer by analysing a set of data from a database known as ShotLink. ShotLink is a platform used by the PGA Tour to collect real time scoring of every shot of every player of every PGA event. While a golf score at the end of a round may indicate what amount of strokes the golfer took to complete the round it will explain the individual factors it took for the individual golfer to achieve that score.

This dissertation will analyse what it takes for a golfer to shoot a successful round by measuring on course statistics by understanding and interpreting the different kinds of golf shots that take place in a round/tournament. Golf is an extremely skilled game with many aspects to always improve on as there is no such thing as a perfect golfer. While most people with a general sports knowledge will have heard of Tiger Woods who dominated the game for several years even he had many flaws statistically, more importantly he was successful in the correct statistical areas. This analysis will examine such areas as Putting, Driving and Chipping, some of the key areas of golf in major depth. It will also look at why some golfers fail and what can help them improve.

I will also be looking at how a golfer can evade the torrid "Cut"; a halfway point in the tournament where players usually outside the top 70 do not feature for the rest of the tournament.

1.2 Motivation

The idea for this project came about as a young struggling amateur golfer and enthusiast, as well as having a keen interest in statistics and numbers. Having an interest in this area will make this project more enjoyable to complete.

I have been fascinated with the levels of technology and engineering that are applied to developing golf equipment, all just aimed at helping to get a small white ball into a hole. It's not just about the use of laser technology or GPS systems but when you see a company like Boeing Aircraft being involved in improving the aerodynamics of golf clubs, you know it must be important.

But of course it's not just about having the best equipment, as I have discovered myself. It's about how and who it's used by. That's where it starts to get interesting, in particular for the top professional golfers. TV stations pour out statistics on golf performance by the players, such as average driving distance, number of putts, sand saves etc. But they rarely seem to get it right, as to who might win or more importantly how they might win based on past performance statistics. This has given me a huge interest in digging more into statistics in golf to see what is available, if it could be used better and to test the statistics.

1.3 Research Problem

This dissertation is being completed as a final year project. When I realised that the project could be completed on a selection of the individuals choosing, my first thought was to approach sports statistics. Since then I have stuck with my initial thought and have focused solely on Golf. Selecting the relevant data was an integral part of this project. After much research, I decided that ShotLink was the best source of information to which I could pull my data from. ShotLink features data from the PGA Tour dating back to when it was introduced in 2005. While ShotLink is a private database that requires access approval, there has been plenty of academic research carried out based on ShotLink's data since its origin for academic purposes. To date the extent of research has been utilized for more specific focus. Many examples include, the implementation of new stats, algorithms that help obtain new data and some general analysis research papers at a basic level of understanding the data. While this research may be intriguing and worthwhile, I still believe there is immense potential for further work that can be done in this area. The majority of research has been completed using simple methods. Through the use of predictive analysis and machine learning models, I hope to make a breakthrough in assisting in the performance of professional golfers.

The main research problem at hand for this dissertation is determining if the analysis methods selected can be applied appropriately, in order to understand and predict trends or patterns in the performance of a professional golfer on the PGA Tour. Another critical component to take into account is, is the retrieved information from ShotLink suitable for answering the proposed questions at hand; does the data allow for different techniques of statistical analysis.

1.4 Challenges

Curse of Dimensionality

There are several challenges that I plan to face in order to complete this project. One of the key challenges being proposed is the data at hand. ShotLink provides an enormous amount of exportable data via their online database platform. ShotLink administer PGA Tour data dating back to 2005 at various different entry levels. This includes datasets detailed by Event, Hole, Round, Radar Trajectory and Radar Launch levels. For the purpose of this dissertation the bulk of the analysis will be completed on the event level data. Each dataset contains relevant data for each year on the PGA Tour, for example the stroke dataset for 2015 would contain approximately 1 million rows and about 40 columns, essentially containing every shot, every player hit, for every tournament of that year. After much consideration, I decided that event level was the best form of the data that could be used for this project as it covered the statistics for each player in each event by year, as well as been smaller in size. The decision process for this selection was helped through the understanding of the "Curse of Dimensionality". The Curse of Dimensionality is a term used in machine learning that according to (Leahy, 2015) is "the more columns you have, the more rows of data you will need in order for that machine learning model to be effective." The event level data has nearly 200 columns. In order to make this data more convenient, I will not be examining every single column available. Reduction techniques will be applied to narrow down the focus of what data is critical to the analysis.

Research

Another important challenge is evaluating previous research. The use of previous research whether it be related to ShotLink or not, will help form a better understanding and provide guidance to the experimentation. With the correct interpretation, I can use the research to my advantage. For research purposes, I will not just be looking at previous Golf Analytical papers, I will be taking all forms of Data Mining into account. This will be explained in further detail in my Literature Review.

Psychology

A considerable challenge to take into consideration while analysing this data is the Psychology of a golfer. This data is purely data on what happens when the golf ball is hit by the golfer, it does not take into account any external or off-course factors which can affect golfers in today's game. This research will be purely statistical.

Results

The most integral part of this project is generating the results. The analysis will be complete using statistical programming languages such as R. The likes of Excel will be used to make the data manageable. Platforms such as Weka and SPSS will also be used for statistical and predictive analysis. This will be the most challenging and time consuming aspect of this project. Significant experiments will be carried out to generate these results, while all may not be successful the projected results will be interpreted within this dissertation.

Interpretation

After the eventual results are generated the key challenge will be interpreting the results. The results will provide key statistical observations on the PGA Tour and attempt to forecast if there is a key formula to success on the PGA Tour. The challenge of interpreting the results will be measured off the initial proposed questions been asked, for example what are the key components in someone winning the Masters?

1.5 Objectives

Research Objectives

The main objective of this project is to apply machine learning techniques as well as further statistical analysis methods to determine and predict what makes a successful golfer on the PGA Tour. To generate such a result, several hypotheses and statistical tests will be proposed about the PGA Tour which will be tested with the ShotLink data. If the projected results are accomplished they are intended to inform and enlighten a professional golfer on how they could approach a specific tournament by measuring factors such as; what did the previous winners of such tournament do well, was there any reasoning between why players missed the cut and made the cut, is there a key approach to shoot a score consistently under par or is there an area of the game that is causing a certain individual to fail, e.g. Putting.

1.6 Methodology

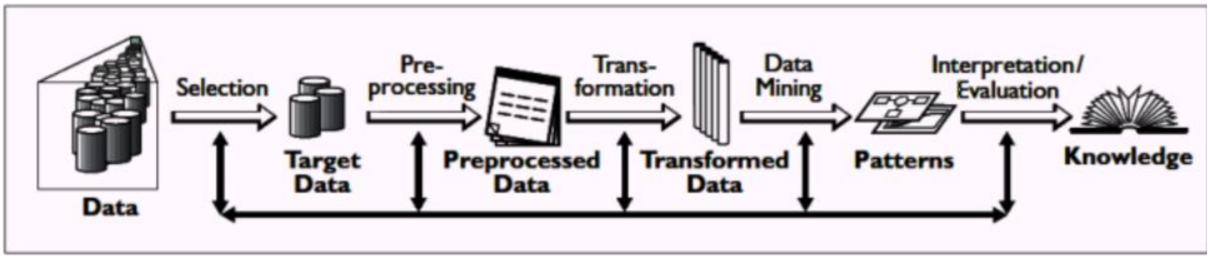


Figure 1 - KDD Methodology

KDD

KDD (Knowledge Discovery in Databases) is a process used in data mining, a methodology I will be applying when implementing machine learning techniques. It provides a high-level overview of the procedures needed to approach a data mining method. While there are other methodologies such as SEMMA (Sample, Explore, Modify, Model and Assess) I believe KKD methodology was best suited to my approach of this project.

Step 1: Selecting the Data

The first step of a data mining process is selecting the data. In this case, the data being selected is that of ShotLink. When selecting the data, it is critical that you understand it. As ShotLink is a database that contains several million records with some non-relevant data, the whole database will not be used. The data was selected with relevance to the suitability of proposed questions / hypotheses and size. Larger datasets would have been more difficult to manage and mine. ShotLink was selected as it had the most thorough information compared to other options such as data provided by "Yahoo Sports". I have narrowed down the selection of data from ShotLink by only taking data from certain years and allowing for certain specifications.

Step 2: Pre-Processing

In my opinion this was one of the most crucial parts of this process. In order to have efficient working data, it first must be pre-processed and cleansed. The removal of noisy data and insignificant outliers can prove to be effective in the long run as the inclusion of these could potentially affect the application of data mining techniques. The removal of null values was a major part of this project. There was a high amount of null values present in the dataset. While the data was not permanently deleted, it was removed in filtering for the appliance of specific queries. It was important to ensure that the data did not become skewed throughout the pre-processing. This process also included the renaming of certain variables and organising data by dates. This was a very time consuming process.

Step 3: Transformation

The transformation stage consists of completing the preparation of the data before the data mining takes place. The data is transformed by reducing the size of the data by removal of non-relevant variables. Throughout the dataset there were several variables that had no correlation to the rest of the data so were therefore removed e.g. “Permanent Tournament Number” and “Team ID”. A PCA (Principal Component Analysis) was completed in order to remove these variables. Such variables were intended for other research purposes. Once these processes had been completed the data was ready to be mined.

Step 4: Data Mining

An appropriate data mining technique must be decided upon. It is important to take into account what method best suits the data at hand and results you are hoping to generate e.g. selecting clustering to represent if subsets of golfers can be divided by “Birdies” and “Greens in Regulation”. Selecting the appropriate technique is vital in producing the best patterns within the data. The generated results of the selected data mining model / technique will produce results of potential interest to this project.

Step 5: Interpretation / Evaluation

Interpreting the mined patterns is the key aspect of this dissertation. The results are the final goal of the project. The interpreted results are the final step of the KDD process which generates the knowledge which, in this case, is how to predict the performance of a professional golfer on the PGA Tour. This dissertation is the knowledge of this process. In order to interpret the results correctly one must have a full understanding of how the data mining process has been applied to the data and what the results mean. To summarise this process, the ShotLink data has been selected, cleansed, transformed, mined and then interpreted to produce the knowledge of interesting patterns that lie within the data. This knowledge attempts to answer the proposed questions been asked prior to the process.

1.7 Resources

In order to successfully complete this project, the following technical and non-technical resources are needed:

- DELL workstation running Windows 10
- ASUS FX553VD

- ShotLink Data
- IBM SPSS Statistics 23
- RStudio
- Internet Access
- Microsoft Word / Excel 2016
- Notepad ++ 7.3.2
- Back Up Storage Services (Google Drive, College Drive)
- Project Supervisor

1.8 Scope & Limitations

Since golf has become a sport that relies heavily on technology in recent years, the research is quite varied with regards to ShotLink. From my research, I have discovered plenty of websites that attempt to predict golf statistics. The majority of these websites were betting related and did not have any machine learning generated results. They were purely taken from a betting perspective and in some cases basic algorithms were applied. In the cases of were data analytic tools were applied, tools such as "SAS Enterprise" and "Rapid Miner" were used. While these are very helpful tools for data mining, they do not generate as detailed results as the likes of a statistical programming language like R. I Will be using similar software within SPSS and WEKA but R will be the main point of focus for this analysis.

As previously mentioned the "Curse of Dimensionality" could potentially be an issue within this project. The sheer size of the might an influence on the success of the model's performance in generating the results of predicting performance of a professional golfer. The chances of this were decreased through the application of dimension reduction in the KDD process.

It is also imperative that I do not create a false representation of ShotLink's data. The results will be based purely on the data that has been provided by ShotLink. Any results that will have been generated through the use of altered or misused data will be insignificant and of testament to ShotLink.

The scope of this project is to assess the performance of professional golfers on the PGA Tour through the means of ShotLink data. It aims to predict trends and patterns in the success of golfers who have previously had success and failure on tour. It will also take into account what it takes to win certain tournaments. The predictive model will inform golfers what events and courses best suit their game and how they can make the most out of their PGA Tour membership, from making the cut to winning majors. This dissertation is aimed at all levels of professional golfers.

For the likes of psychological matters regarding golfers there may not be any statistical data to prove whether or not certain psychological factors can affect a golfer's performance. An attempt will be made to prove particular hypotheses such as "Does a golfer's round go downhill after a triple bogey?" These hypotheses will be tested using the ShotLink data. There will be no external datasets to be used in this project apart from ShotLink's.

1.9 Golf Explained

Golf is a game that is said to have been invented in the early 18th century in Scotland but did not become popular until the late 19th century. I will now provide a brief explanation of how the game is played and the basic formatting. There are plenty of frequently asked questions among non-golfing people that are asked.

Golf is a game played outdoors. A golf course normally consists of 18 holes. To complete these 18 holes a golfer must have a set of no more than 14 golf clubs which they will use to hit the ball into the hole. Typically, a golfer will have a Driver to hit the ball off the tee, a set of Irons for approaching the greens, and a putter for putting the ball into the hole. This is a standard setup for any golfer, it gets a little more complex for the professionals.

The notation of "Par" is a concept that many people struggle to understand. Even I myself had trouble understanding it when first taking up golf. Each hole will have a set score of par; 3, 4 or 5. These are usually measured by the distance of the hole. To complete a hole in par a player must get the ball in the hole the required number of shots also known as strokes. For example, if a player gets the ball in the hole on a par 3 in three strokes they have made "Par". If they take more than three strokes they have finished the hole "Over Par" and if in less than three strokes they have completed the hole "Under Par". An average golf course usually spans around Par 72 meaning you would have to take 72 shots to complete the round level par. This would be considered standard for a pro but highly unlikely for a weekend golfer.

*See Appendix for full Overview of Golf Terms.

2. Requirements Specification

2.1 Introduction

This Requirements Specification document is being designed with the intention to of outlining the requirements to be set out for Assessing Professional Golfer Performance on PGA Tour Data Analytics Research Project. This document aims to outline the overall description of the document and all the specific requirements needed including any constraints or limitations.

2.2 Purpose

The Purpose of this Requirements Specification is to inform those who have a chartered interest in obtaining ShotLink data. The document intends to inform the management team of the “ShotLink” Database of the specific requirements. This document’s intended audience and users are those who wish to gain a further understanding of the requirements at hand.

2.3 Description

This section of the document will provide a perspective for the project and all the functions to be carried out. Functional and Non-Functional Requirements will be discussed. It will also highlight any assumptions or dependencies concerning the database as well as from a user’s perspective.

2.4 Constraints or Limitations

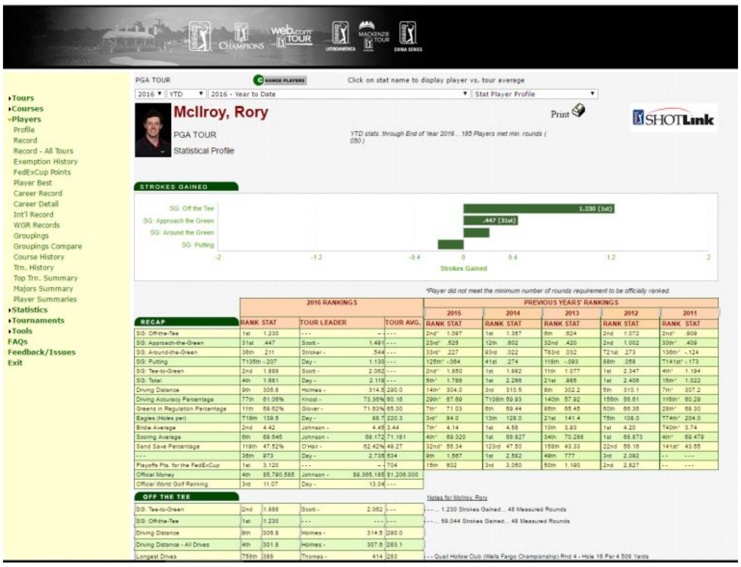
Specific Data must be requested from the Data Stewards if needed, for example if the pre-generated exported data is a detailed request that can be generated to query more specific results.

ShotLink consists of a GUI which features a capacity of information not accessible to download as viewed in the image below.

Connection to “ShotLink” Database requires a user login and password from a successful generated request from “ShotLink” Data Stewards.

Larger data files may be time consuming if not specified correctly.

The Results are dependent on “ShotLink” Database maintaining consisting data.



- Tours
- Courses
- Players
  - Profile
  - Record
  - Record - All Tours
  - Exemption History
  - FedExCup Points
  - Player Best
  - Career Record
  - Career Detail
  - Int'l Record
  - WGR Records
  - Groupings
  - Groupings Compare
  - Course History
  - Trn. History
  - Top Trn. Summary
  - Majors Summary
  - Player Summaries
- Statistics
- Tournaments
- Tools
- FAQs
- Feedback/Issues
- Exit

PGA TOUR

Click on stat name to display player vs. tour average

2016 | YTD | 2016 - Year to Date | Stat Player Profile

McIlroy, Rory

PGA TOUR

Statistical Profile

YTD stats. through End of Year 2016... 185 Players met min. rounds (050)

STROKES GAINED

*Player did not meet the minimum number of rounds requirement to be officially ranked.

	2016 RANKINGS				PREVIOUS YEARS' RANKINGS									
					2015		2014		2013		2012		2011	
RECAP	RANK	STAT	TOUR LEADER	TOUR AVG.	RANK	STAT	RANK	STAT	RANK	STAT	RANK	STAT	RANK	STAT
SG: Off-the-Tee	1st	1.230	--- --	---	2nd*	1.097	1st	1.367	6th	.624	2nd	1.072	2nd*	.909
SG: Approach-the-Green	31st	.447	Scott - 1.491	---	23rd*	.525	12th	.602	32nd	.420	2nd	1.002	30th*	.409
SG: Around-the-Green	38th	.211	Stricker - .544	---	33rd*	.227	93rd	.022	T83rd	.032	T21st	.273	138th*	-.124
SG: Putting	T135th	-.207	Day - 1.130	---	126th*	-.084	41st	.274	118th	-.093	88th	.058	T141st*	-.173
SG: Tee-to-Green	2nd	1.888	Scott - 2.062	---	2nd*	1.850	1st	1.992	11th	1.077	1st	2.347	4th*	1.194
SG: Total	4th	1.681	Day - 2.118	---	5th*	1.786	1st	2.266	21st	.985	1st	2.406	15th*	1.022
Driving Distance	9th	306.8	Holmes - 314.5	290.0	14th*	304.0	3rd	310.5	8th	302.2	5th	310.1	7th*	307.2
Driving Accuracy Percentage	77th	61.06%	Knost - 73.36%	60.16	29th*	67.69	T108th	59.93	140th	57.92	156th	56.61	116th*	60.29
Greens in Regulation Percentage	11th	69.62%	Glover - 71.63%	65.00	7th*	71.03	6th	69.44	66th	65.45	60th	66.36	28th*	68.30
Eagles (Holes per)	T19th	139.5	Day - 88.7	220.3	3rd*	84.0	13th	128.0	21st	141.4	T5th	108.0	T74th*	204.0
Birdie Average	2nd	4.42	Johnson - 4.45	3.44	7th*	4.14	1st	4.58	10th	3.93	1st	4.20	T40th*	3.74
Scoring Average	6th	69.646	Johnson - 69.172	71.181	4th*	69.320	1st	68.827	34th	70.298	1st	68.873	4th*	69.479
Sand Save Percentage	119th	47.52%	O'Hair - 62.42%	49.27	32nd*	56.34	123rd	47.50	158th	43.33	22nd	56.16	141st*	43.65
---	36th	973	Day - 2,735	534	9th	1,567	1st	2,582	49th	777	3rd	2,092	--	---
Playoffs Pts. for the FedExCup	1st	3,120	--- --	704	15th	602	3rd	3,060	50th	1,190	2nd	2,827	--	---
Official Money	4th	$5,790,585	Johnson - $9,365,185	$1,206,000										
Official World Golf Ranking	3rd	11.07	Day - 13.04	---										

OFF THE TEE					Notes for McIlroy, Rory
SG: Tee-to-Green	2nd	1.888	Scott - 2.062	---	--- ... 1.230 Strokes Gained... 48 Measured Rounds
SG: Off-the-Tee	1st	1.230	--- ---	---	--- ... 59.044 Strokes Gained... 48 Measured Rounds
Driving Distance	9th	306.8	Holmes - 314.5	290.0	
Driving Distance - All Drives	4th	301.8	Holmes - 307.6	283.1	
Longest Drives	T58th	388	Thomas - 414	283	--- Quail Hollow Club (Wells Fargo Championship) Rnd 4 - Hole 16 Par 4 508 Yards

Figure 2 - Screenshot of ShotLink GUI

2.5 Assumptions and Dependencies

It is assumed that:

All the exported data will be correct and up to date.

The user will not create a false representation of the data in “ShotLink.”

The user will not violate any of the pre-agreed Terms & Conditions.

The user has a basic understanding of “ShotLink” Fundamentals.

The user will be able to identify any problems or issues that arise within the Database and notify the data stewards.

That only authorized personnel will have administrative access to the database.

Dependencies:

There will be a working internet connection.

Full functionality of the required Software and Technologies.

Access to the required resources needed.

2.6 Requirements Specification

2.6.1 Functional Requirements

The following functional requirements are what is needed in order for the user to successfully obtain the required data needed for the projected results for the project. For each circumstance the referred "User" is myself.

2.6.1.1 Functional Requirement 1.1

Identification Code:

FR1

Title:

Access to ShotLink

Description:

User must request access to "ShotLink" Database through ShotLink website: http://www.pgatour.com/stats/shotlinkintelligence.html.

Priority:

Essential (High Priority)

2.6.1.2 Functional Requirement 1.2

Identification Code:

FR2

Title:

User ID and Password Login

Description:

User will receive a Unique User ID and Password Login, which will allow them to enter "ShotLink" Database.

Priority:

Essential (High Priority)

2.6.1.3 Functional Requirement 1.3

Identification Code:

FR3

Title:

Accessing ShotLink

Description:

User will enter their Unique User ID and Password via stats.pgatourhq.com. User will accept Terms & Conditions and will be granted access to “ShotLink.”

Priority:

Essential (High Priority)

2.6.1.4 Functional Requirement 1.4

Identification Code:

FR4

Title:

Analysing ShotLink Interface

Description:

Once User is presented with main page they will have the option to select from “Tours”, “Courses”, “Players”, “Statistics”, “Tournaments”, “Tools”, “FAQ’s”, “Feedback/Issues” and “Exit”. The user has the option to access all of these features.

Priority:

Essential (High Priority)

2.6.1.5 Functional Requirement 1.5

Identification Code:

FR5

Title:

Exporting Relevant Data

Description:

Once the user has familiarised themselves with the ShotLink Interface, they will be expected to export the data needed. They will have several options to export data. The user will select “Tools” > “Detail Export”. The User will then be able to select from the following Export levels: “Stroke Level”, “Hole Level”, “Round Level”, “Event Level”, “Course Level”, “Radar Launch”, Radar Trajectory” and “Better Worse Same”. Once the User selects an export level they will have the option to download pre-generated export(s) based upon year. The data will be downloaded in a .txt file. The user will also have the option to generate one ‘filtered’ export based on year, tournament, and other criteria in the same export level. The user will enter their specific request along with their email address details.

Priority:

Essential (High Priority)

2.6.1.6 Functional Requirement 1.6

Identification Code:

FR6

Title:

Requesting Relevant Data

Description:

The user will also have the option to request specific data if the pre-generated data is not tailored to their needs. In this case the user must select “Tools” > “Statistical Analysis Tool”. Once selected, the User will have the option to select two players and compare statistics for different tournaments and rounds. The user will receive the specific data once the Data Stewards approved their request. They will be notified by email with the requested data.

Priority:

Not Essential (Moderate Priority)

2.6.1.7 Functional Requirement 1.7

Identification Code:

FR7

Title:

Downloaded Data

Description:

Once the user has retrieved the relevant data, the following steps should be followed as seen in figure 3 in order to successfully manage and analyse data in the required Software / Analytical tools.

Priority:

Essential (High Priority)

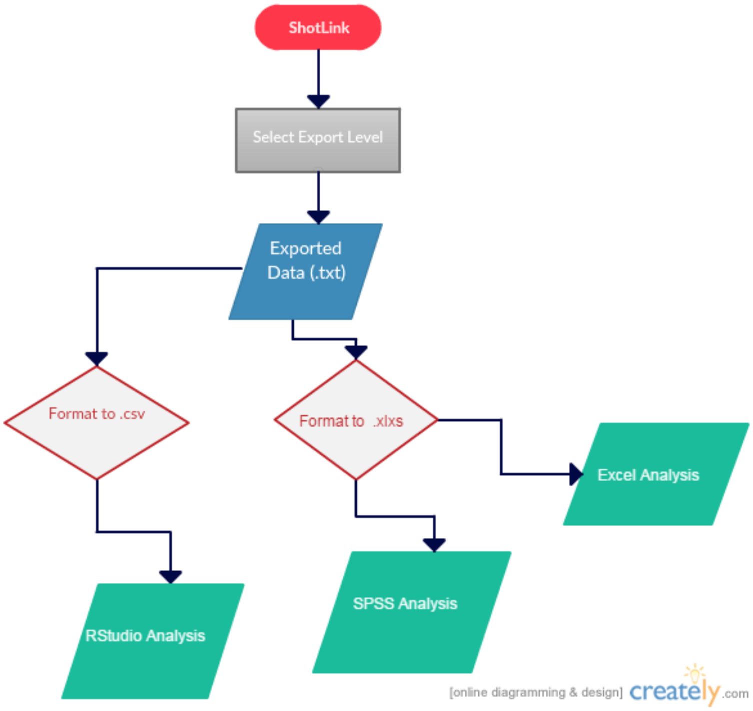


Figure 3 - Data Flowchart

2.6.2 Non-Functional Requirements

2.6.2.1 Security Requirements

Identification Code:

NFR1

Title:

Terms & Conditions

Description:

User must agree to maintain agreement with Terms & Conditions agreed with “ShotLink” upon approval of access.

2.6.2.2 Performance Requirements

Identification Code:

NFR2

Title:

Response Time

Description:

It is required that ShotLink has a consistent Response Time when returning results of its GUI.

2.6.2.3 Availability & Reliability Requirements

Identification Code:

NFR3

Title:

Reliability

Description:

ShotLink relies on real time data of live golfing events. Potential cancellation or possible unforeseen circumstances during events may affect the consistency of the data.

Identification Code:

NFR4

Title:

Available Data

Description:

ShotLink must maintain data on a consistent and regular basis. Any requested data by a User must be made available to such user in a reasonable period of time. If data requested by the user is unavailable, the data stewards shall inform such user of this occurrence.

2.6.2.4 Recovery Requirements

Identification Code:

NFR5

Title:

Recovery

Description:

If Data is misplaced in ShotLink, it is expected that the data is recovered as soon as possible.

2.6.2.5 Capacity & Scalability requirements

Identification Code:

NFR6

Title:

File Size

Description:

All exported data should be of rational file size. The User should have adequate space available on their Personal System.

2.6.2.6 Maintainability Requirements

Identification Code:

NFR7

Title:

Database Maintenance

Description:

It is assumed that ShotLink Data Stewards will maintain consistent data and updates to ShotLink.

2.6.2.7 Usability requirements

Identification Code:

NFR8

Title:

User Friendly

Description:

ShotLink should commit to preserving a User-Friendly GUI. If changes are made to ShotLink the user will be notified.

3. Analysis Techniques & Data Interpretation

3.1 Data Mining and Statistical Analysis Techniques

With such a large selection of methods to be used for analysing data nowadays, I will briefly discuss some of the methods I am considering to use. It is important to acknowledge whether the data is labelled or unlabelled when deciding on what test to perform.

3.1.1 Supervised and Unsupervised Learning

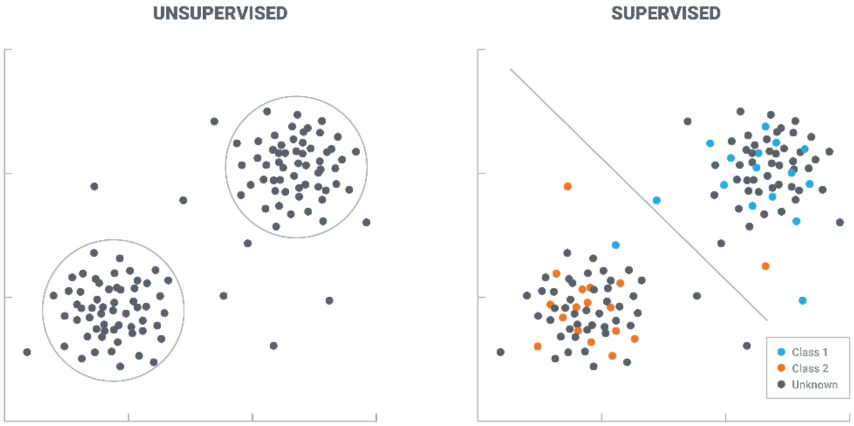


Figure 4 –Unsupervised vs. Supervised

When building a machine learning model the data will fall under one of two classifications, labelled or unlabelled. This will indicate whether the model will be of supervised learning or unsupervised learning. Labelled data is when the data has knowledge or meaning to it, unlabelled data is data with no significance or meaning. For example, a tweet is just a tweet and is considered unlabelled until value is applied to it. For example, the data for a tweet can be considered labelled if sentimental value is applied to it.

Supervised learning applies to the majority of machine learning models. A supervised learning model learns the data from an already labelled trained dataset to produce an output. The goal is essentially to input an x value and return a y value with some sort of predictive knowledge. I will mainly be using supervised learning models for this project.

Unsupervised learning uses unlabelled data. It has an x input but no targeted y value of an output. It attempts to learn from the train data to create new classes or clusters, to provide information and value.

3.1.2 Multiple Linear Regression

Multiple Linear Regression is a model built on two or more variables. Every value of x (the independent variable) is associated with a value of y (the dependant variables), a linear relationship between a set of dependant and independent variables. Simple Linear Regression is used when there are two continuous variables, one been the predicted (dependant) variable. If two variables are highly correlated a Simple Linear model can be built based on those variables data. A golfing example would be predicting the Number of Birdies a player would make based on the Number of Greens in Regulation they hit based on previous data. Birdies and Greens in Regulation have 82.1% correlation; determined using Pearson's Correlation Coefficient, this would infer that a strong reliable model would be built. When creating a Multiple Linear Regression Model the possibility of Multicollinearity must be tested for, if there are independent variables that are highly correlated to each other. If Multicollinearity is present an independent variable must be removed.

3.1.3 Clustering

Clustering is a process that uses unsupervised learning. Cluster analysis is used to group a set of data into subsets or clusters. Each group / cluster is assigned by similarity; each variable / object will have relevance to its cluster. For example, a cluster may contain golfers with higher average driving distances and another cluster may contain golfers with lower average driving distances. This will be a method I will be using frequently for visualisation and analysis purposes. See figure 5 for example.

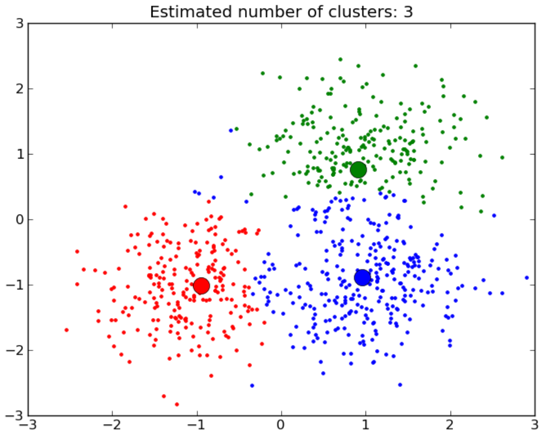


Figure 5 - Clustering Example

3.1.4 T-Tests

A statistical hypothesis test that determines if there is a significant difference between two sets of data. A null and alternate hypothesis is stated. After testing for Normality, a decision must be made on whether a Paired or Unpaired T-Test is to be carried out. Unpaired T-Test compare two different subjects. Paired T-Tests makes comparisons between the same objects.

3.1.5 KNN (K Nearest Neighbour)

KNN is similar to how a decision tree works. It is a “Lazy Learner” Classification and Regression method. KNN is known as a “Lazy Learner” as it does very little work when training the model. KNN methods perform the majority of their classifying / processing when predicting the actual classification. As per (Han et al., 2012) “nearest neighbours classifiers learn by comparing test data with training data which is similar”. Classification is done by searching for nearest neighbours in the data, it assumes that observations close to each other have an impulse to belong to the same observation.

3.2 Understanding the Data

Pre-processing and Transforming

One of the key aspects prior to starting the analysis was understanding the data in full. As previously mentioned cleansing and transforming the data was critical to having a dataset that could be analysed. This turned out to be a time-consuming process due to the large volume of data present.

As the exported data from “ShotLink” was in .txt format it had to be converted into. xlxs and .csv format. The text data was also disordered so a delimiter was used when importing the data into Microsoft Excel to format variables by commas and tabs. This process was repeated for each data file.

One of the main issues when analysing this data was “The CUT” data, because players are removed from tournament at the “Cut Line” after round 2. This meant that several variables contained null values such as “Money” and “Round 3 & 4 Score”. The majority of variables where a statistic had a rank value for the 4 rounds meant that when a player missed the cut they had a value of “999” instead of a true value of 5th (5) for example in “GIR Rank”. When analysis was being performed in RStudio these values were removed when applying functions, e.g. when calculating the average Birdies made by players who made the cut, the function would apply (! =999) value not equal to 999. This retained all the values not equal to 999 or in other words all the players who missed the cut. I also created a column with values “Yes” and “No” to which if a player made the Cut. Another issue for some variables was that did they not necessarily reflect on a player’s performance. For example, with “Total Holes

Over Par" one player who missed the cut may have the same amount of Holes over Par as someone who made the cut. With one player playing 36 holes and the other playing 72 holes. To provide a more accurate value I have depicted each relevant variable as a percentage. This was done using Microsoft Excel by applying formulas to each row. An example would be dividing "Total Greens in Regulation" by "Total Holes Played" and applying the function to a new row called "GIR%". This is a commonly used statistic but was not included in this dataset. These percentages showed more clearly that players who missed the cut had a lower percentage of GIR than players who made the cut.

Another issue when preparing the data was outliers / noisy data. After performing a Principal Components Analysis in SPSS there were several irrelevant variables to the analysis. Such variables included "Round 5 & 6 Score", "Team ID" and "Official Event". Such variables had no significant influence on the rest of the data. Some data such as players withdrawing during tournaments "W/D" and "DNS" did not start influenced the analysis, so for that purpose they were removed. This did not skew the data as these instances were lowly populated within the data.

When examining the data in RStudio most variables had integer values (int). In some cases, these variables were converted into Factor values. "Player Age" is one example that was changed to Factor. This was for purposes of plotting graphs.

During first analysis of the data in RStudio I noticed that when searching for certain results extra criteria needed to be added to the queries to retain specific results. I decided that sorting the data through excel would be more efficient in the long run. Example, sorting Finish Position by lowest number to highest number. This allowed for easier viewing of the data in RStudio. This also made it easier to cluster the data into sufficient groups of similarities.

3.3 Hypothesis Testing

A hypothesis test is a form of statistical testing that is based on a set of sample data to speculate whether that specific condition is true for an entire population. A set of hypotheses must be set out before the tests are carried out. A Null and Alternative hypothesis are set out. The null hypothesis is the statement that is been tested. Commonly the null hypothesis is that there is no difference between the sample data. The alternative hypothesis statement infers the opposite that there is a difference. In most cases, we are hoping that the null hypothesis is true. The statements for each hypothesis can vary depending on the type of hypothesis test been carried out.

Before deciding on a form of Hypothesis test a normality test must be carried out. This test measures if data is normally / ordinally distributed or not normally distributed. It determines the "goodness of fit” of a normal model to the data. We know if the data is normal if the tested p value is equal to or above 0.05 of whatever the alpha value is that has been chosen for that test. If the data is normally distributed, we can carry on with whatever hypothesis test we originally decided on. If the data falls below the alpha value (significance level) we must consider a different hypothesis test. There are a set of specific hypothesis test for non-normally distributed data known as Non-Parametric Tests.

3.4 Proposed Questions

Is age a factor for professional golfers in the modern game?

What variables influence success on the PGA Tour?

Can a Golfer overcome the dreaded Cut?

Do Golfers achieve more success from Distance or Accuracy?

What separates the Best from the Rest?

Did Tiger Woods’ Marriage Scandal affect his performance?

Is it possible to Predict Tournament Results for Individual Players?

4. Research

4.1 Literature Review

Assessing Golfer Performance

In recent years, players have become more reliant on the use of statistics to inform them about their performance. I will now discuss some of the statistical terms used in golf and evaluate some of the research that has been carried out in this department.

"Is it possible to assess Golfer Performance through the use of various statistical and data mining techniques?"

Introduction

Prior studies reveal that in recent times the application of statistics in golf has become a critical aspect in measuring and enhancing the performance of a professional golfer. Most successful studies have been completed using ShotLink data.

Previous Work

Many papers to date have solely focused on the "shot value" aspect of research. For example, (Broadie 2012) used Shot data provided by ShotLink to measure performance of professional golfers. The shots were assigned by 5 categories; tee, green, fairway, rough and sand. Broadie used a model to measure course difficulty and the skill of the player. Factors in course difficulty were measured by distance and scoring averages. The model assumed that player's skills were fixed and did not change at every changing event, this predictive model does not allow for a multiple of changing factors over time. Broadie produces a metric that measures "Strokes gained to the field" per individual player. While this metric is successful it does not consider that the level of golfer skill varies throughout each tournament, a skill that cannot be measured with ShotLink data.

As per (Stockl et al. 2013) the ISOPAR method was introduced. "The ISOPAR method is used for characterizing the difficulty of golf holes and allows the performance of shots to be analysed." This incorporates ball location data provided by ShotLink. The analysis intends to incorporate a value that one shot has after the one before. Once again the author does not consolidate that golfer quality varies over time.

(Leahy, 2015) also incorporates ShotLink data to measure and predict golfer performance. The focus of this analysis is solely on how can a golfer better themselves to make the cut. Leahy uses predictive statistical methods such as clustering to build models that train the ShotLink data to generate whether an individual will make the cut or not. Although this model is useful it does not consider some factors that could influence whether or not a professional golfer would attempt to use this knowledge.

(Riccio Ph.d 2012) Produces a metric that determines which player is the best ball striker from the fairway on the PGA Tour. Riccio won the “ShotLink Intelligence Prize” for his work. The model measures shots from 150 and 225 yards to exclude wedge shots and purely measure that of full swing shots taken from the fairway. This model suggests which players are best at striking the ball from the fairway. The results indicate that on average 80% of greens are hit from 150 yards and 40% of greens are hit from 210 yards. This clearly supports that theory that the further a player is from the hole the harder it is to get close to it. In my opinion a compensation to be made from this research is that the shorter hitters off the tee would require to be better ball strikers from the fairway to make the most of their game, this would apply to players like Zach Johnson.

One of more noteworthy pieces of research was also provided by (Broadie, 2008). Through the research from measuring the performance of amateur golfers Broadie helped to create a new statistic in concurrence with work of (Fearing et al., 2010) called “Strokes Gained – Putting”. The statistic is now used by the PGA Tour and is included in ShotLink’s data. It argues the fact that when previously calculating a player’s putting stats for a round, the total number of putts was not a significant validation of how good a putter they are, because it doesn’t measure the length of the putts. The algorithm for strokes gained putting measures the mean number of it takes putts it takes to hole out and the definite number of putts it actually takes.

Data Analytics in Other Sports

In modern times, data has become an ever-important factor in generating results and statistics. From betting sites to leading management consulting companies, data analytics has been implemented at some level.

A more recent example is Accenture, who used predictive analytics to forecast match scores in the RBS Rugby Six Nations. The data consisted of match related scoring data (80%) and sentimental analysis of live Tweets (20%). As seen below in figure 6 Accenture measures the win probability against live statistics.

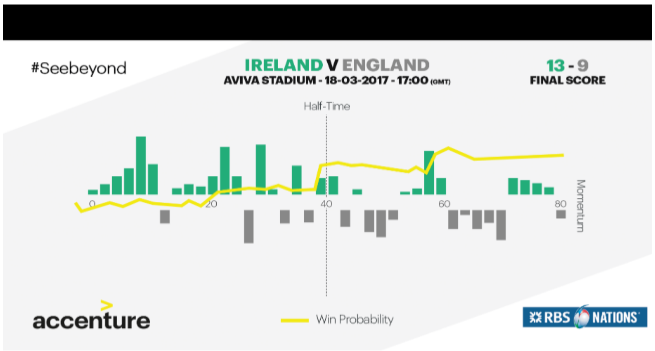


Figure 6 - Accenture RBS Six Nations Predictions

One of the first examples of using statistics in sport for decision-making purposes was “Moneyball: The Art of Winning an Unfair Game” a book written by Michael Lewis in 2003 (Lewis, 2003). Lewis used Sabermetrics (empirical analysis of Baseball) to interpret the likeliness of teams winning and number of runs in a match. A movie was released in 2011 based on the book, “Moneyball”.

In some cases, entire websites are dedicated to research in sports science and statistics and the relation they have to predicting performance in sports. “FiveThirtyEight.com” is a website that was created by “Nate Silver” in 2008. Silver’s website proved to be a great success through his use of statistical and data mining techniques in predicting sports performance. Silver’s methods proved to be so useful they were later applied to politics. His analytics system was eventually used to predict the outcome of how 49 of 50 states voted in the Presidential election that year.

As well as predicting political results, data analytics is used in many different areas, from forecasting weather, identifying retail habits, measuring housing prices or even predicting locations of future outbreaks for infectious diseases. These are just some examples of how data is been used to provide meaningful information in the modern era.

Hypothesis Testing in Sports

Hypothesis Tests are commonly used to test for statistical differences in Sport. (Clemens, 2017) examined if there was a difference in the performance of Roger Clemens, a professional baseball player before and after he was tested positive for the use of performance enhancing drugs. A T-Test (paired) was performed on the data prior to drug use and after. The results suggested that there was no statistical evidence to suggest that his performance had changed before and after the drug use.

(Verma, J. 2016) used hypothesis testing to measure the flexibility of three professional gymnasts. He used a Kruskall-Wallis test (A non-parametric One Way Anova) to test if there was a difference between the flexibility of the group of three gymnasts, whether they were equal in flexibility or different. After (Verma, J. 2016) performed the Kruskall-Wallis test he concluded that there was no difference in the flexibility of the three professional gymnasts.

Conclusion

In conclusion, it is clearly evident that there is more than one way to analyse golf statistics in today’s environment. The importance of ShotLink to this research is also critical which shows its importance to the world of golf. A common occurrence when assessing golfer performance is the variability of changing components such as golfer quality. This is why there has been no major breakthrough in this area. Although these are more likely for psychological purposes it is difficult to predict golfer performance if you were to measure performance by changing quality of each individual golfer and not fixed variables. Research also shows that complex algorithms and complicated machine learning models are not completely necessary when generating results. Studies from the likes of (Broadie, 2008) show that simple approaches can be taken to retrieve significant results from ShotLink data. Along with Golf, it is clear the affect that Data Science and Analytics has had on other Sports in recent times. Research suggests that the use of data will only develop how information is produced in Sports in the coming years.

5. Data Analysis

5.1 Exploring the Data

Once the data has been cleansed and prepared successfully we can begin an exploratory analysis. This analysis will be complete in RStudio and SPSS.

5.2 Descriptive Statistics

By generating descriptives in RStudio we can gain a deeper understanding of the data. At first look we can see that there is (5547) observations and (190) variables in the dataset. By applying the (str) function we can see what data type each variable is. For this dataset Integer and Factor will be the main data types used. When first examining the dataset I noticed the bulk of variables were influenced by “The Cut”. This made it very difficult to analyse the data due to the high number of “999” values. There was also an inconsistency in the data as for some variables such as “Birdies” there was variance in the data as the players who made the cut and played four rounds would have more Birdies than the players who missed the cut and played two rounds. 58% of the data consists of (3235) players who made the cut for those events. The other 42% consists of (2313) players who did not make the cut. There were 44 events that took place in 2016. Due to the fact of inconsistencies in some variables I have created two new data frames on top of the current overall event data. The data has been grouped into players who made the cut and players who missed the cut. This allows for easier calculation of certain statistics / variables.

5.3 Money

While most golfers play for their passion of the sport it is their job at the end of the day. Money is the main incentive for most individuals whether they choose to admit it or not. Making the Cut is every golfers goal after the first two rounds of every tournament, every player is fearful of missing the dreaded Cut and not being paid for that week. This is why it is so difficult for some players to make it on Tour, as missing consecutive cuts can be very costly after all other expenses. In 2016, the average earnings on the PGA Tour was $95,517 per event. Jason Day was the highest earner in an individual event earning 1,800,900 winning the Player’s Championship. The four major winners were also among the highest earning individual events. The lowest winning earnings was $6,060 for Dawie Van der Walt in the Puerto Rico Open. The Puerto Rice Open and the Barbasol Championship had the lowest mean earnings of tournaments in 2016.In 2015 Jordan Spieth made $12,030,465 in prize money alone, excluding sponsorship deals. Not bad for playing a game you enjoy. Players at the midpoint of the money list usually earn about 600,000 to 700,000 dollars per year on the PGA Tour. Throughout this dissertation.

5.4 Age

Golf is one of the sports that essentially has no retirement age. Although it is a sport that features high levels of athleticism it is not required. The average age on Tour at the moment is 33. Youth seems to be taking over the game at the moment with the youngest winners last year only 22 years of age; Jordan Spieth and Justin Thomas. Henrik Stenson was the oldest winner at 40 years of age. The mean age of winners was 31.

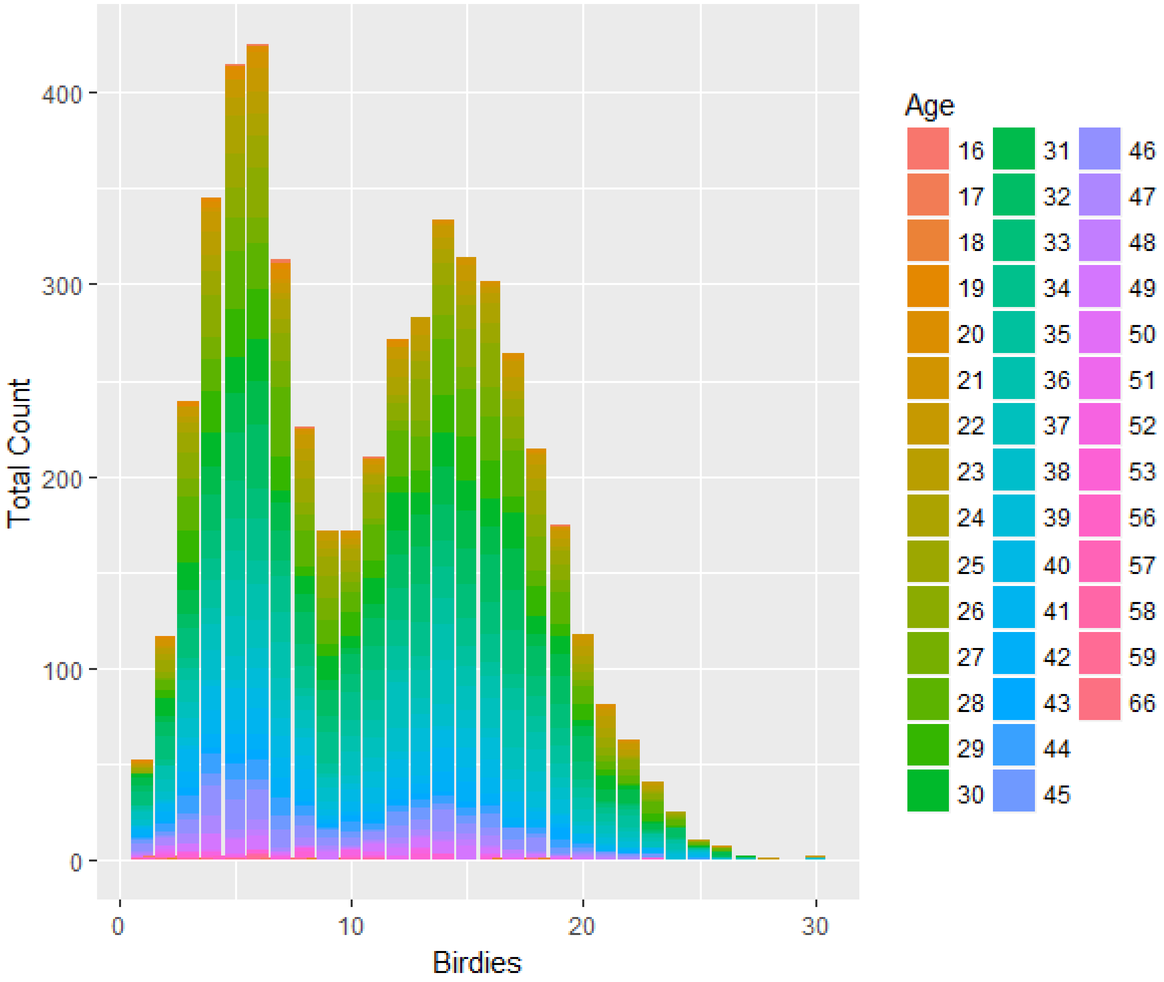


Figure 7 - Histogram of Birdies to Age

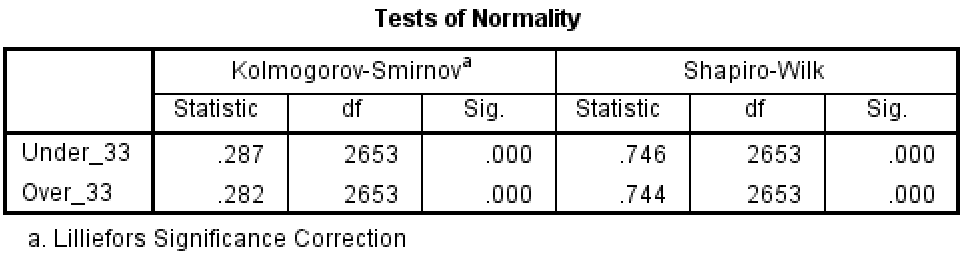
From the above figure 7 we can see the influence age has on the number of Birdies made in an event. The histogram shows that players over 40 usually fall within the lower range of birdies made in one week. Players ranging from mid-20's to mid-30's occupy most of the birdies made in 2016 events.

5.4.1 Testing Age

A hypothesis has been set out to test if there is a difference in the success between a group of golfers under the age of 33 and golfers 33 and over.

We must first test for normality to see what type of test is to be carried out. If the data is non-normally distributed a non-parametric test is to be carried out. After testing for normality we have reported a p-value of <0.001. This is less than the significance value of 0.05 which means the data is non-normally distributed and a non-parametric test must be carried out.

Table 1 - Normality Test for Age Data



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Under_33	.287	2653	.000	.746	2653	.000
Over_33	.282	2653	.000	.744	2653	.000

a. Lilliefors Significance Correction

Wilcoxon Signed Rank Test

A Wilcoxon Signed Rank Test will be used for this experiment

Null Hypothesis – H0: M (<33) = M (>33)

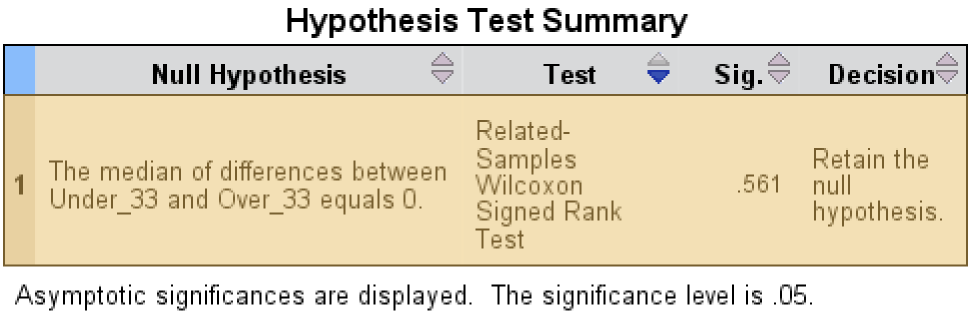
There is no difference in the medians between Finishing Positions for Golfers under 33 and Finish Positions for Golfers 33 and Over.

Alternative Hypothesis – H1: M (<33) ≠ M (>33)

There is a difference between the medians of Finishing Positions for Golfers under 33 and Golfers 33 and Over.

After performing a Wilcoxon Signed Rank Test on the sample data we can conclude that we will retain the Null Hypothesis that there is no difference in the medians between Finish Positions for Golfers under 33 and Finish Positions for Golfers 33 and Over after reporting a p-value of .561.

Table 2 - Wilcoxon Test Result



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between Under_33 and Over_33 equals 0.	Related-Samples Wilcoxon Signed Rank Test	.561	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

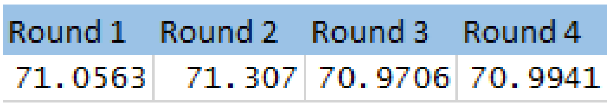
This result suggests that although younger players are said to be the best in the world this evidence supports the theory that there is no difference in the performance of age on the PGA Tour. There is slight evidence to support that younger players earn more money as the average earnings for players under 33 years of age was $100,155 compared to the average earnings of over 33's of $90,390. Quite a significant difference considering there was no difference in the medians of Finish Positions between these two sample groups.

5.5 Round Scoring

One of the most critical aspects to be successful in Golf is having good Round Scores. At the end of the week the winner of the Tournament is whoever has the lowest number of Total Strokes or the lowest accumulative score to Par after four rounds. I will be analysing if there are any trends in whether the scoring of each round is influential to the final Total Strokes score or the Cut mark.

At first look from the boxplot in figure 8 we can see that each Round Score has the same median of 71. As seen below Round 2 has the highest average score with Round 3 having the lowest. These results might explain a few things. Firstly, round 3 average score might explain the term moving day; when a player attempts to shoot a low round to make a move on the rest of the field. Round 2 average score might be the highest due to players losing focus after knowing they are not going to make the cut. One reason why Rounds 3 and 4 are lower on average than 1 and 2 might be because the better players make the cut which means the scoring average should be lower.

Table 3 - Average Score of Round Level



Round 1	Round 2	Round 3	Round 4
71.0563	71.307	70.9706	70.9941

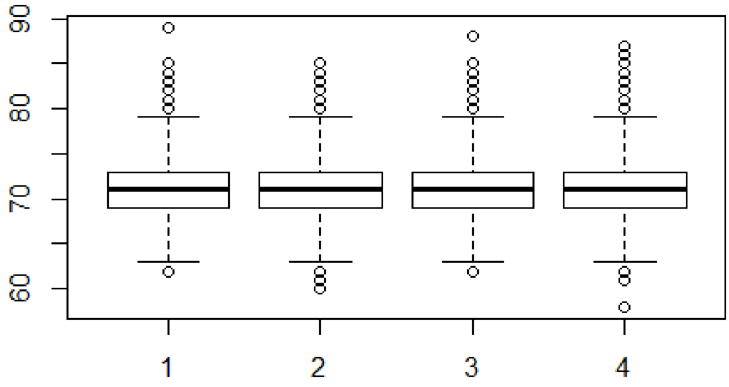
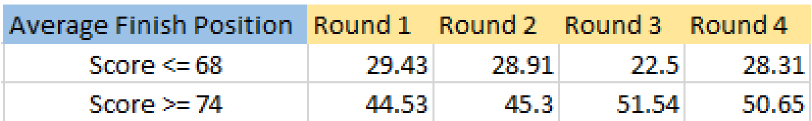


Figure 8- Boxplot of Round Scores 1-4

The lowest round score was 58 by Jim Furyk in Round 4 of the Travellers Championship who ended up finishing 5th. He was in 70th position after the end of round 3. He was 14 shots better off than his round 3 score. That week 70th Position earned $13,068 where Jim Furyk earned a significantly higher $231,825. This shows the difference a low Round score can make especially the final round. The highest round of 2016 was 89 in round 1 of the US Open by Zachary Edmondson. The highest opening round score for a player who made the cut was 83 by Kristoffer Broberg in the Cadillac Championship. He shot a second round 64 allowing him to make the cut and eventually finish 64th.

Another question to be asked is "is there a round / day to have a best score on?"
For the purpose of this test let us say a significantly good score for one round is 68 (3 shots lower than the average score) or lower.
Over the first two rounds for every event in 2016 there were 2240 players who shot 68 or lower in either round 1 or 2. 90% of these players made the cut where 10% of players who shot a 68 or better missed the cut. When we analyse players who shot a 74 or worse in either round 1 or 2 we can see that 80% of those players missed the cut, compared to the 20% of players who made the cut after shooting 74 or worse.

Table 4 - Average Finish Position when shooting <68 or >74



Average Finish Position	Round 1	Round 2	Round 3	Round 4
Score <= 68	29.43	28.91	22.5	28.31
Score >= 74	44.53	45.3	51.54	50.65

From table 4 above we can see the influence of scoring in each round. We can see that the best day to shoot a good score is on Saturday (Round 3) the average finish position for players who shoot a 68 or lower in Round 3 is significantly lower than any other round. This could prove the theory of “moving day”, that Saturday is the most influential day in regards to where the player will finish, with the assumption that the majority of golfers will not have a disastrous final round. Interestingly enough when we apply the same criteria to players who shot a round of 74 or worse we see that round 3 and 4 are much higher than 1 and 2. This shows the importance of bad scores at the weekend; note that the average finish position regardless of any scoring is approximately 36.

5.6 Round 1 Leader

Of the 81 Round 1 leaders (including ties) 2 missed the cut. The average finish position for first round leaders is 18th. On the PGA Tour 18th position guarantees a lot of money. There were 10 occasions where the first round leader went on to win. We can see from figure 9 that most Round 1 leaders end up finishing in the Top 10.

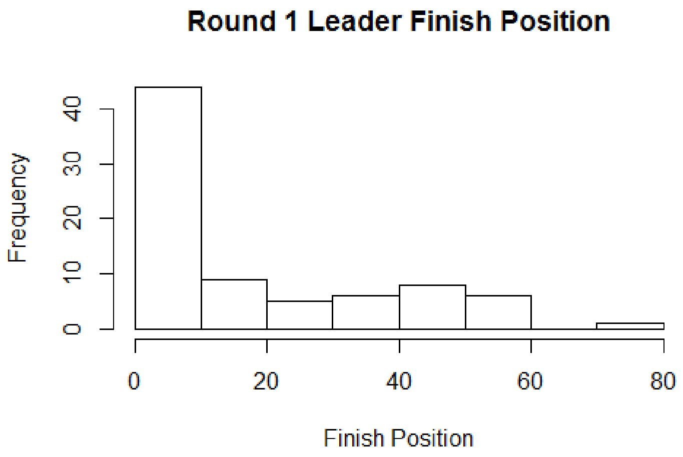


Figure 9 - Finish Position of Round 1 Leaders

5.7 The Cut

The elusive cut is one of the most dreaded scenarios for every professional golfer on tour. This section will analyse the key aspects of the Cut.

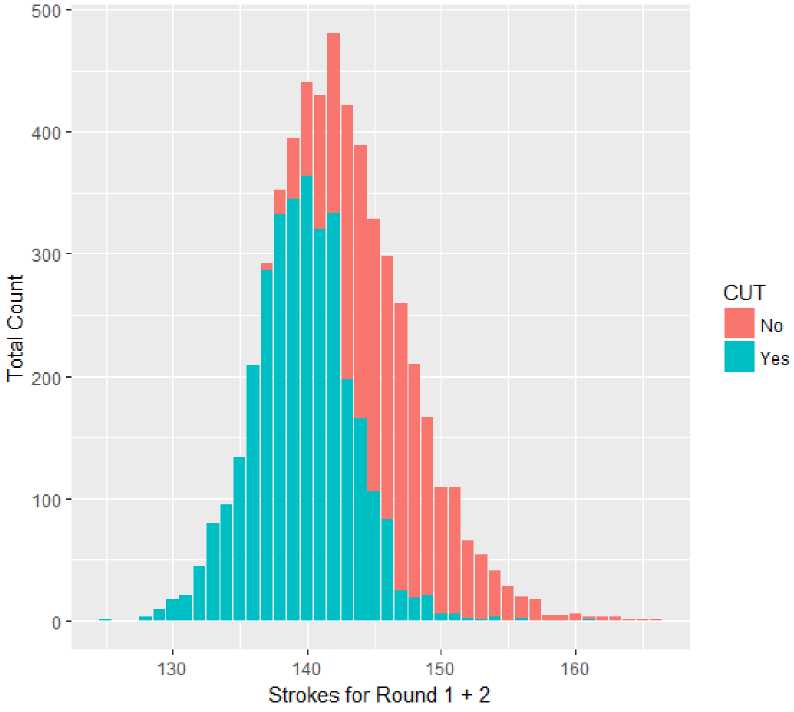


Figure 10 - Stroke Count for Cut

To provide a fair adjustment of the Cut mark I have created a new variable "Round 1 + 2 Strokes"; the total strokes a player completed round 1 and 2 or their combined scores for round 1 and 2. We could not provide an accurate depiction of scoring terms from the "Total Strokes" variable as players who made the cut had much higher total strokes of four rounds than the two rounds of the players who missed the cut. The above graph figure 10 shows the relevance of scoring for the first two rounds. The average cut mark is approximately around 145 strokes per tournament. A score of 73 in round 1 and 73 in round 2 for a player would indicate that they would miss the cut on that given week. There were two interesting instances where a player had a "Round 1 +2 Strokes" Value of 137 and still missed the cut. This was Derek Fathauer and Spencer Levin in the "Career Builder Challenge". This may have been because the par value for that course was considerably lower than other courses or that the scoring for that week was very low across the field.

6 Key Aspects of Scoring / Success

In Golf the question everybody wants answered is how to shoot a good score? Is it by making lots of birdies, avoiding bogeys, hitting the ball further, been more accurate or putting well? These attributes along with many more all add up to the total stroke count at the end of the week. The attributes have been selected from analysing previous research and creating a correlation matrix to see which attributes are highly correlated to scoring factors such as "Total Strokes" and "Finish Position". I will now analyse all of these attributes to see what does in fact help a golfer be "successful on the PGA Tour". After reviewing what attributes contribute to success I hope to then forecast if player performance can be predicted using these attributes. Finish Position has been divided into the following clusters, figure 11.

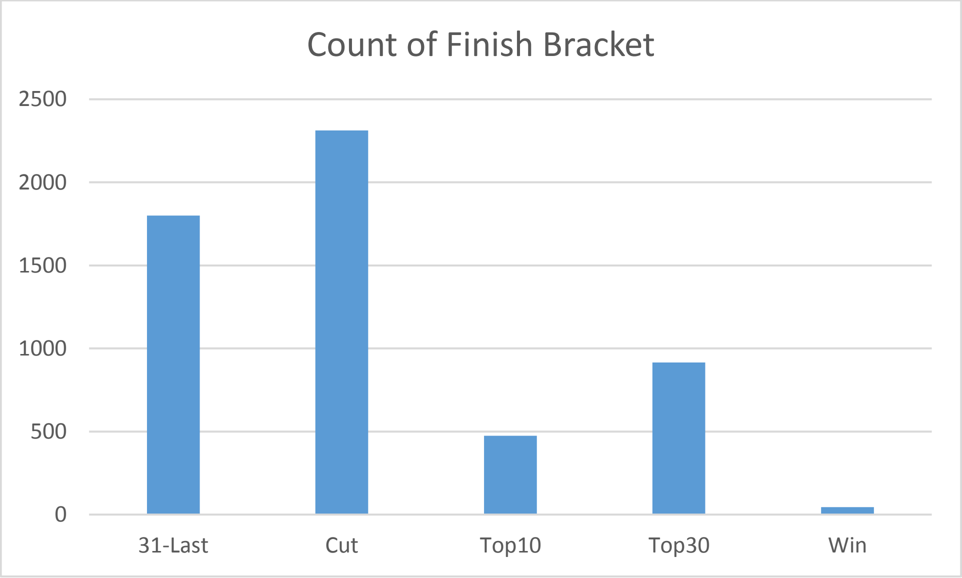


Figure 11 - Histogram Count of Finish Bracket Positions

6.1 Scoring

From figure 12 we can see the importance of each scoring term with pars been the obvious factor. More birdies are made than bogeys while slightly more eagles are made than others (triple bogeys or worse).

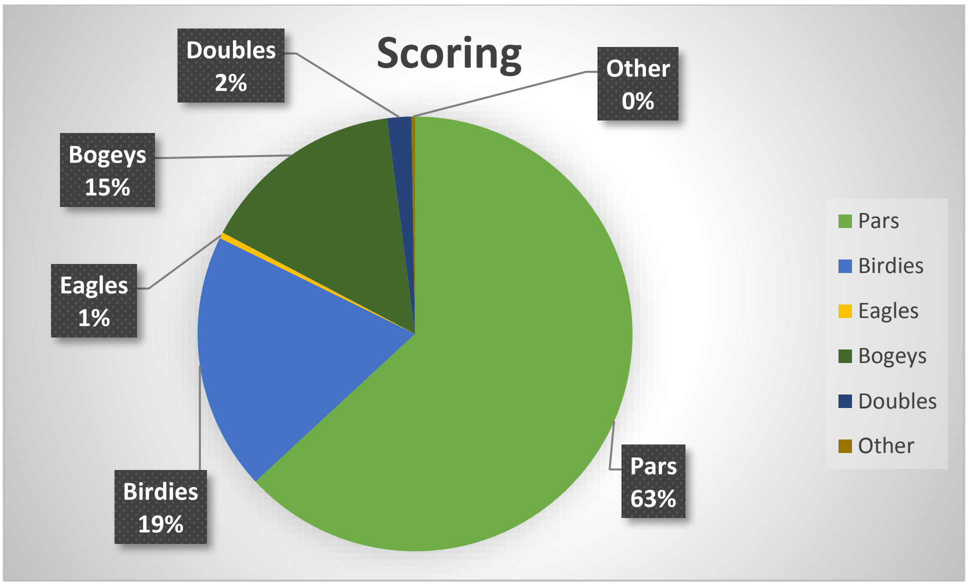


Figure 12 - Pie Chart Representing Scoring Terms

6.2 Birdies / Bogeys

As we know making Birdies is the key for any golfer in scoring under par. Figure 13 illustrates the significance of Birdies and Bogeys to where a player will finish. From the table 5 below we can see the top 11 players for birdie averages.

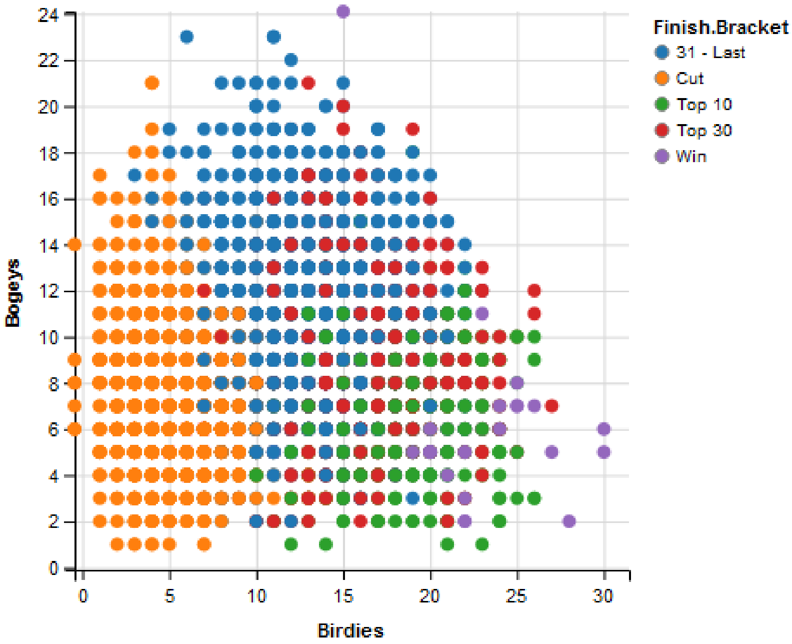
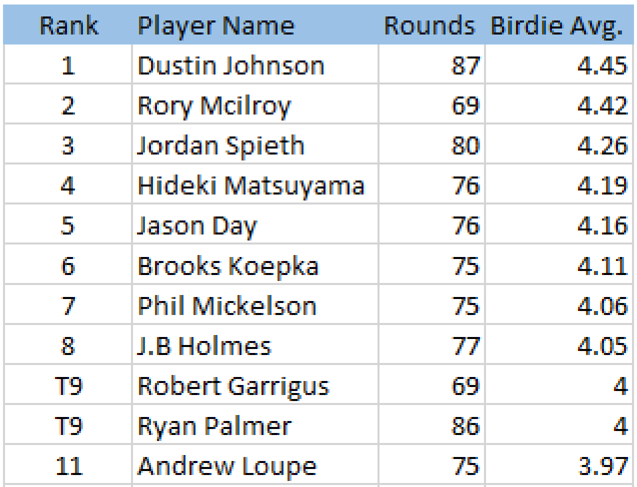


Figure 13 - Scatterplot Birdies to Bogeys clustered by Finish Bracket

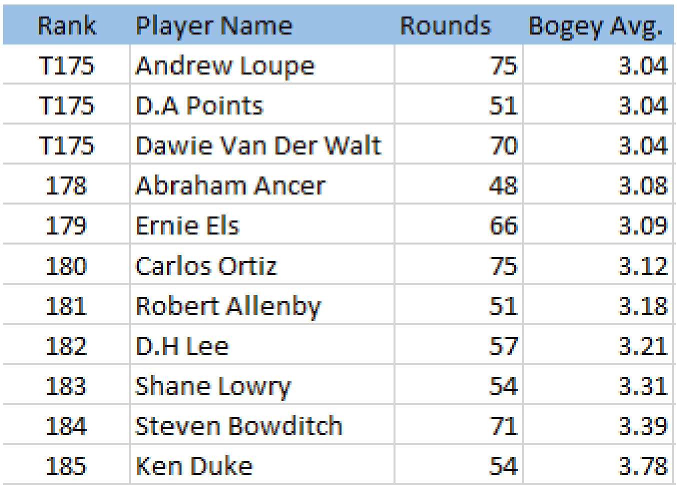
Table 5 - Top Birdie Avg.



Rank	Player Name	Rounds	Birdie Avg.
1	Dustin Johnson	87	4.45
2	Rory Mcilroy	69	4.42
3	Jordan Spieth	80	4.26
4	Hideki Matsuyama	76	4.19
5	Jason Day	76	4.16
6	Brooks Koepka	75	4.11
7	Phil Mickelson	75	4.06
8	J.B Holmes	77	4.05
T9	Robert Garrigus	69	4
T9	Ryan Palmer	86	4
11	Andrew Loupe	75	3.97

Two interesting names in this table are Robert Garrigus and Andrew Loupe. Loupe is the only player on this list to have not won on the PGA Tour and Garrigus has only one once, back in 2010. The other 9 players are multiple winners on Tour.

Table 6 - Bottom Bogey Avg.

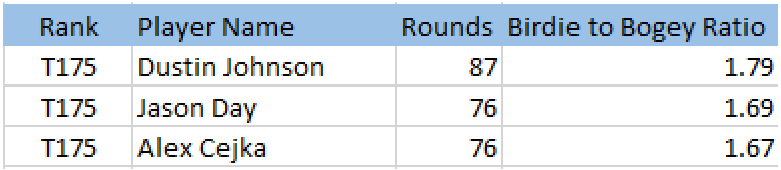


Rank	Player Name	Rounds	Bogey Avg.
T175	Andrew Loupe	75	3.04
T175	D.A Points	51	3.04
T175	Dawie Van Der Walt	70	3.04
178	Abraham Ancer	48	3.08
179	Ernie Els	66	3.09
180	Carlos Ortiz	75	3.12
181	Robert Allenby	51	3.18
182	D.H Lee	57	3.21
183	Shane Lowry	54	3.31
184	Steven Bowditch	71	3.39
185	Ken Duke	54	3.78

The above table 6 explains Andrew Loupe's performance on the PGA Tour for 2016. The table illustrates the Bogey Average per player. Andrew Loupe is near the bottom with 3.04 bogeys on average per round. Rickie Fowler was 1st with 2.03, A hole bogey less on average.

We can also see why Dustin Johnson was the most successful player on Tour in 2016. He was 1st in Bi rdie avg. with 4.45 and fifth in Bogey avg. at 2.21. This brings us on to Birdie / Bogey ratio.

Table 7 - Top 3 Birdie to Bogey Ratio



Rank	Player Name	Rounds	Birdie to Bogey Ratio
T175	Dustin Johnson	87	1.79
T175	Jason Day	76	1.69
T175	Alex Cejka	76	1.67

As we can see from table 7 Dustin Johnson as expected clearly leads the way for number of Birdies to Bogeys along with Jason Day in 2nd; two of the best player in the world. Alex Cejka's inclusion is quite intriguing however. Throughout 2016 he played 19 events, 4 of which he finished in the Top 10 and 6 where he missed the cut. Considering there was very little difference between the Birdie to Bogey Ratio, you would think that they earned similar prize money for that season. Jason Day earned $6,425,110 compared to Alex Cejka's $1,559,362. This may be simply because Jason Day was more efficient in the tournaments where more money was at stake, but it should be examined none the less.

At first observation, we can see that nearly 64% of Jason Day's earnings were in 3 tournaments alone, 1st,1st and 2nd, in 3 significant tournaments on the Season calendar. Jason Day only missed one cut out of 17 events, which shows the importance of consistency with regards to money earnings. Notably the one event Day missed the cut in; "The Farmers Insurance Open", Cejka also the missed the cut. Jason Day is considerably younger at 29 compared to Cejka at 46.

From analysing the data, we can see where the main differences lie between the two players. Day was ranked 1st in Strokes Gained Putting, Cejka ranked 92nd. Driving distance was also key to Day's success, he averaged 304.3 yards off the tee compared to Cejka's 281.5. Distance is known to be an important factor in terms of scoring on Tour. Cejka made only 5 eagles to Day's 17. Accuracy is also key to good scoring. Cejka's accuracy rank of 22nd probably explains his high Birdie to Bogey ratio. Day has a low Driving Accuracy, ranked at 181st. This asks the question "Is there a difference in success between the top ranked players for Driving Distance and Driving Accuracy?" We will look at this in further detail later.

Scrambling may also explain Cejka's Birdie / Bogey ratio. He was ranked in 3rd in Scrambling, successfully scrambling Par 65.03% of the time compared to Day who scrambled Par 61.19% of the time. Another attribute that highly contributes to scoring is Greens in Regulation. From the boxplot below in figure 14 we can see that Day and Cejka had similar Green in Regulation percentages throughout 2016. We can make an inference from these results that Jason Day is more of power player who makes a lot of Birdies compared to Alex Cejka who is more of finesse player and bases his game on avoidin g Bogeys. We can see the main difference between the two players comes down to Putting with Day' s stats far superior. We also assume that Day although he is younger has more experience and comp osure in the more important tournaments than Alex Cejka.

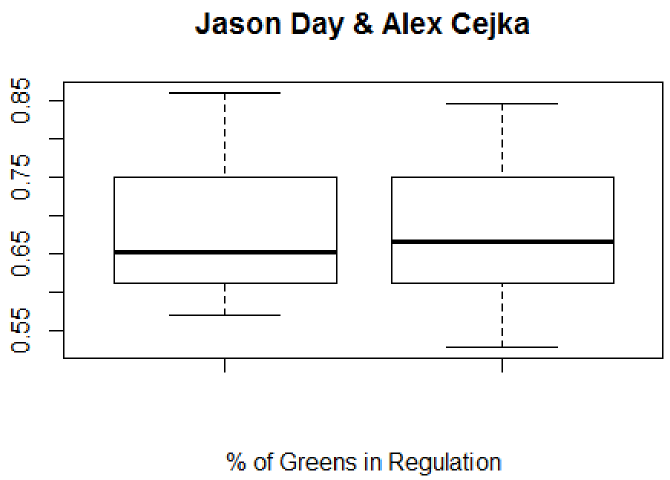


Figure 14 - Boxplot of Jason Day & Alex Cejka GIR% 2016

6.3 Distance vs. Accuracy

Distance and Accuracy are two elements of golf that most players aim to always improve on. We have seen how relevant it is for the likes of Jason Day and Alex Cejka. So how does Accuracy and Distance affect scoring, let's examine it in a wider context.

Driving Accuracy

A test has been set out to see if there is a statistical difference in the Finishing Position of the players who rank inside the top 35 of Driving Accuracy and players who rank outside the top 35 in Driving Accuracy.

After performing a Shapiro-Wilk test we can confirm that the sample data is non-normally distributed.

Wilcoxon Signed Rank Test

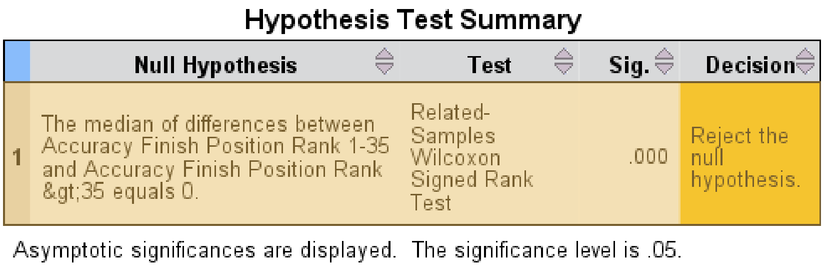
Null Hypothesis – H0: M (Accuracy 1-35) = M (Accuracy 35+)

There is no difference in the medians between Finish Positions for Golfers who rank inside the top 35 for Driving Accuracy and those who rank 35th and above.

Alternative Hypothesis – H1: M (Accuracy 1-35) ≠ M (Accuracy 35+)

There is a difference between the medians of Finish Position for Golfers who rank inside the top 35 for Driving Accuracy and those who rank 35th and above.

Table 8 - Wilcoxon Rank Test Result for Accuracy Test



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between Accuracy Finish Position Rank 1-35 and Accuracy Finish Position Rank &gt;35 equals 0.	Related-Samples Wilcoxon Signed Rank Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

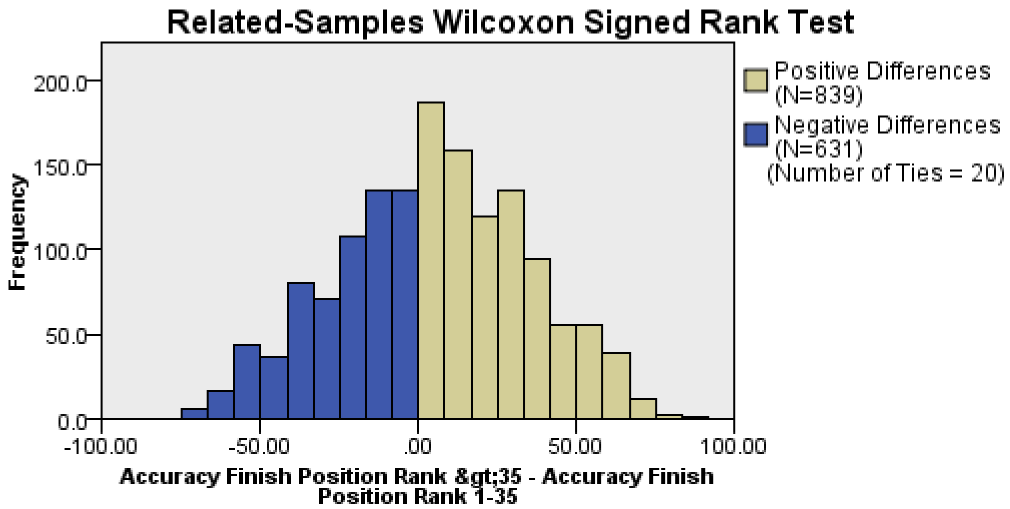


Figure 15 - Histogram of Top Ranked Driving Accuracy Player vs Bottom Ranked

After performing a Wilcoxon Signed Rank Test in SPSS we can report that the Null Hypothesis will be rejected due to the p-value falling below the significance value of 0.05. In favour we accept the Alternative Hypothesis meaning there is a statistically significant difference between the medians of Finish Position for Golfers who rank inside the top 35 for Driving Accuracy and those who rank 35th and above. After performing some descriptive statistics in SPSS we can infer that in fact Driving Accuracy does contribute to good scoring and Finish Positions. Players who rank inside the top 35 for driving accuracy have a median finish position of 31 compared to a median of 36 for players who rank outside the top 35 in driving accuracy.

Driving Distance

We already know that driving distance is critical to performance on the PGA Tour after analysing Jason Day's success. Another example is Dustin Johnson who was 2nd in Overall Driving Distance and 1st in Money Earned per Event; $425,690.

The same test will be carried out on this set of sample data as it is non-normally distributed as well.

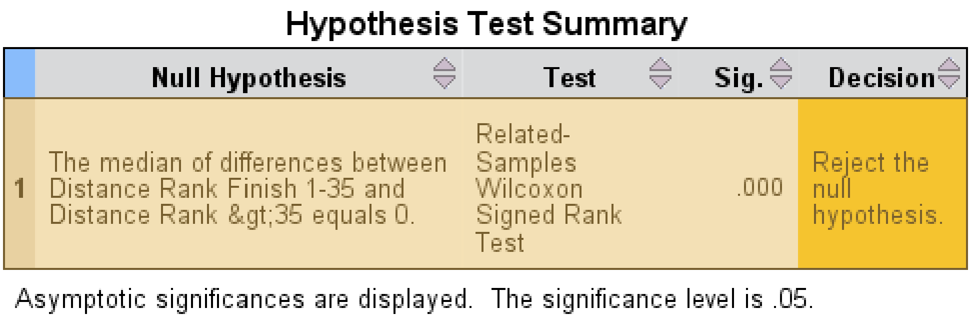
Null Hypothesis – H0: M (Distance 1-35) = M (Distance 35+)

There is no difference in the medians between Finish Positions for Golfers who rank inside the top 35 for Driving Distance and those who rank 35th and above.

Alternative Hypothesis – H1: M (Distance 1-35) ≠ M (Distance 35+)

There is a difference between the medians of Finish Position for Golfers who rank inside the top 35 for Driving Distance and those who rank 35th and above.

Table 9 - Wilcoxon Rank Test Result for Distance Test



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between Distance Rank Finish 1-35 and Distance Rank &gt;35 equals 0.	Related-Samples Wilcoxon Signed Rank Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

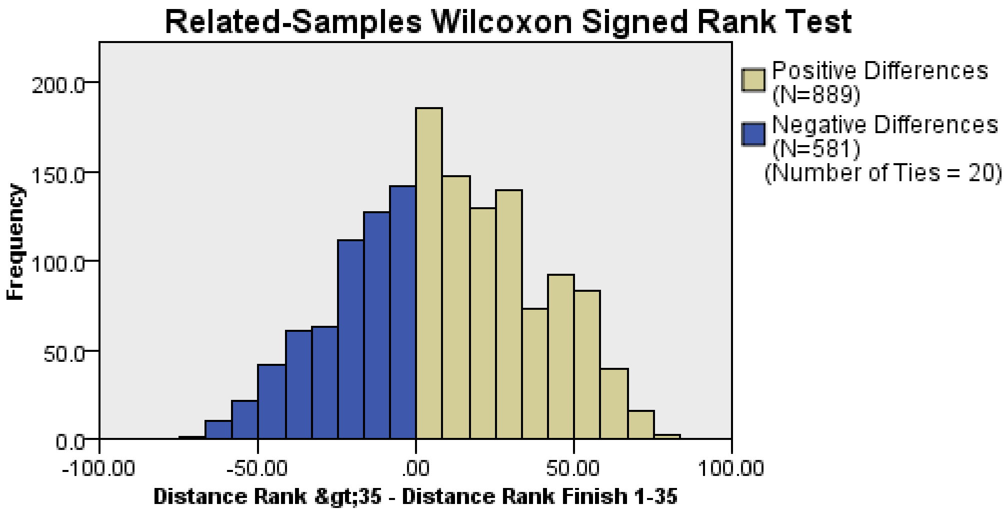


Figure 16 - Histogram of Top Ranked Driving Distance Players vs Bottom Ranked

The results are similar to that of Driving Accuracy test. We are rejecting the Null Hypothesis as the p – value is lower than the significance level of 0.05. We therefore have to accept the Alternative Hypothesis that there is a difference between the Finishing Positions of players who are ranked inside the top 35 for Driving Accuracy and those who are not.

Descriptive Statistics inform us that there is a median finish position of 27 for players who rank inside the Top 35 for Driving Accuracy and a median of 39 for those who did not rank inside the top 35 for driving accuracy.

There is already some evidence to suggest that Distance is more important than Accuracy off the tee, Distance has a lower median and mean finish position then Accuracy. Let's take a further look at whether there is more evidence to support this theory.

Is there a difference in Distance & Accuracy?

For the purpose if this experiment let's take the top 10 ranked players in both categories to see if there is any difference of their performance. While this is examining the two attributes at the higher end of the scale it should still supply enough evidence to answer if either Accuracy or Distance have differences with regards to success. It is also worth highlighting that no player was included in both of these categories. There are 237 and 239 observations for Driving Accuracy and Distance respectively.

Firstly, let's examine which categorical group earned more money that season.

Players who featured in the top 10 for Driving Accuracy earned $105,506.8 on average throughout th e season compared to a somewhat higher figure of $173,362.5 for players ranked inside the top 10 fo r Driving Distance.

Already it is looking as if there is more importance in been longer off the tee than accurate. If we exa mine the results for the two groups we can see a difference. Surprisingly, the players who rank inside the top 10 for Distance make a slightly higher percentage (70.7%) of Cuts than those who rank top 10 for Accuracy (68.7%).

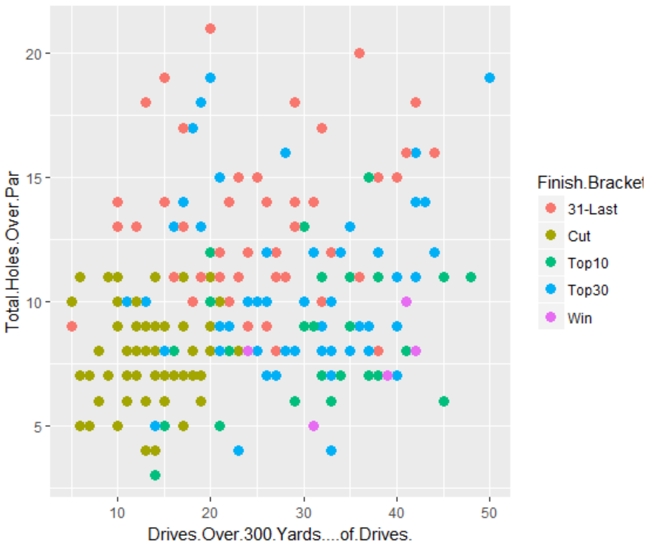


Figure 17 - Scatterplot of Drives Over 300 yards vs Total Holes Over Par clustered by Finish Bracket

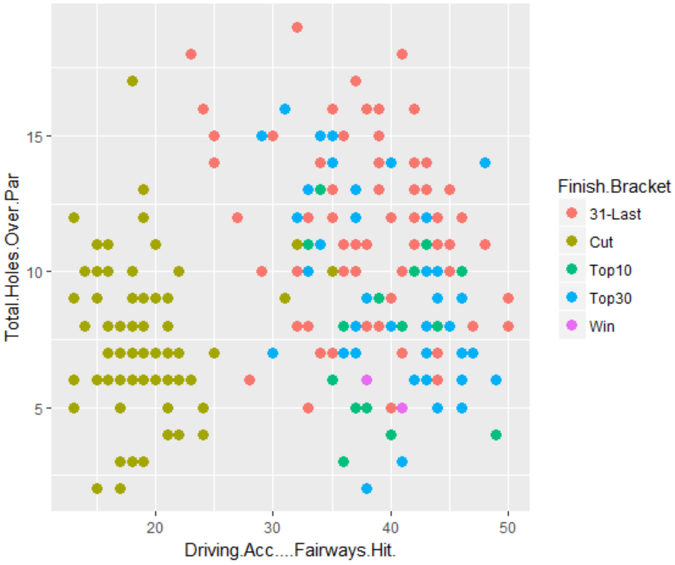


Figure 18 - Scatterplot of Fairways Hit vs Total Holes Over Par clustered by Finish Bracket

Players who fall into the distance category have over twice as many wins as those in the Accuracy category, 7 wins to 3. The players who hit the ball further have also amassed twice as many top 10 Finishes, 40 to 20 respectively. The above plots figure 17 & 18 illustrate the number of Holes over Par for every time a player hit their tee shot over 300 yards and every time a player hit a Fairway. It is evident that there is more consistency in the data for the amount of Fairways hit and Cuts made. In most instances players who hit under 25 Fairways generally miss the cut. We can see that there are several occasions where a player hit less than 25 drives over 300 yards and still managed to be successful, including a win and multiple top 10's. It also illustrates that hitting 30 drives are more over 300 yards and only having 10 or less instances where they were over par can be a guarantee for success, with these stats there are four wins and only two occasions where a player finished 31st or worse. The figure 19 below demonstrates the importance of distance with regards to birdies. The graph highlights that more birdies are made by players who feature in the top ranks for driving distance compared to those who feature in the top ranks for driving accuracy. The x-axis represents the count for the samples; the y-axis represents the number of Birdies made in one tournament by that individual.

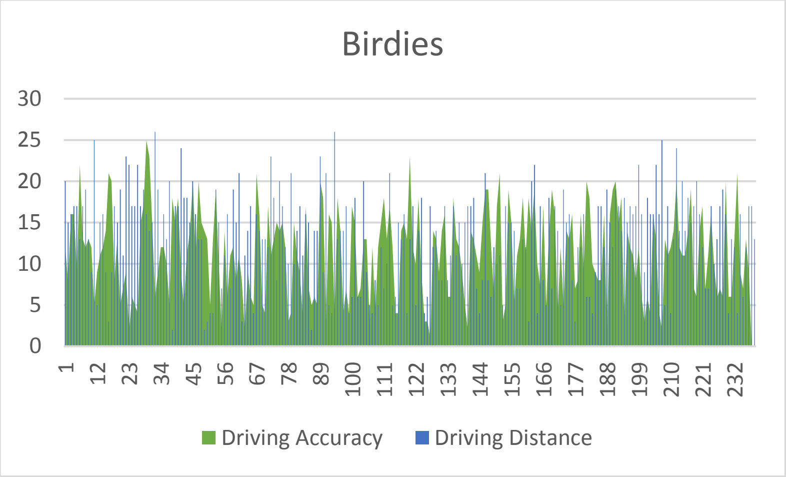


Figure 19 - Sample Birdies for Top Ranked Distance Players vs Top Ranked Accuracy Players

6.4 Greens in Regulation (GIR)

The ability to hit Greens in regulation is said to be a fundamental skill in achieving success in Professional Golf. It is said that each missed green costs a professional golfer about 0.6 strokes, compared to 0.3 strokes of missing a fairway. ShotLink data provides the number of Greens in Regulation hit for each player in each Tournament, whether it be over two rounds or four rounds. I have created a new column GIR% (Total Holes Played / Greens in Regulation). It measures the percentage of Greens in Regulation per player regardless of whether they made the cut or not. Figure 20 represents the significance of Greens in Regulation with regards to the Cut mark.

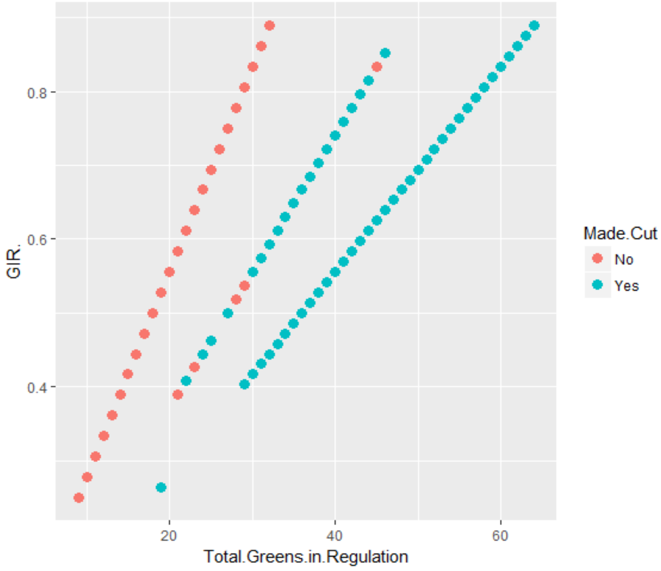


Figure 20 - Correlation of Total GIR VS GIR% measured by Cut

The mean GIR % for 2016 was 64.03%. There are a few interesting observations. Firstly, we can see that one player made the cut with a GIR % of 25%. This is quite the outlier considering no other player in the 2016 season made the cut with a GIR % of less than 40%. The player in this circumstance was Steven Bowditch who eventually finished 58th. One potential explanation might be that Bowditch was extremely successful in scrambling over the first two rounds or he may have "holed out" from when he missed a GIR. Bowditch's putting stats suggest that he holed more putts from outside of 15ft than any other player that week.

On the other spectrum of the scale we can see that there were numerous incidents where a player achieved a GIR % of over 80% but failed to make the cut, this would have been more than likely due to failing to take advantage of these opportunities due to poor putting. Another compelling examination is that the highest GIR % of the season was 89%, this player (J.J Henry) missed the cut. The reason why Henry missed the cut was due to his poor putting. Of the 32 Greens he hit in Regulation he made 5 Birdies which consisted of 4 one-putts. In comparison of James Hahn who also hit 89% of GIR, but made 23 Birdies and 25 One-Putts. He eventually finished 6th. Although he played twice as many holes as Henry it still shows the significant effect Putting has once a Green in Regulation is achieved. One inference we can make is that if a player hits under 40% of GIR they will more than likely miss the cut.

Below in figure 21 & 22 we can observe the relevance of Greens in Regulation to scoring. There is a very high correlation between Greens in Regulation and Birdies (85.11%). As we can see there is also a relatively strong correlation between missed Greens in Regulation and Bogeys (63.07%). The difference in 22% can more than likely be explained by how good professional golfers are at scrambling.

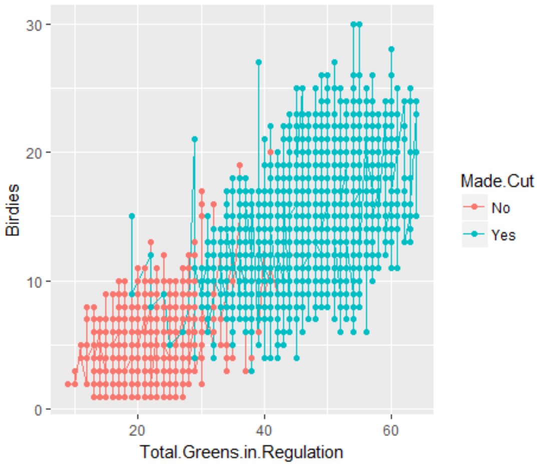


Figure 21 - GIR vs Birdies

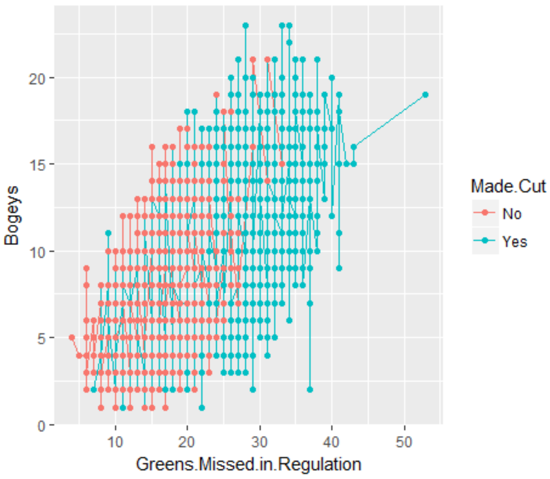


Figure 22 - Missed GIR vs Birdies

6.5 Putting

Putting is reported to account for approximately 40% of all golf shots hit (Gwyn & Patch, 1993). Thus emphasising the importance of Putting to success in the modern game of golf.

By measuring the top performing Putters on Tour last season against their success, we can see that Putting is one of the most significant attributes a golfer can possess in accomplishing success on the PGA Tour. Out of the top 10 players with the best putting average on Tour in 2016, 6 players won at least one tournament. A professional golfer has on average 29 putts in one round of 18 holes. They average around 7 one-putts and .5 three putts per round. Around 74% of Golfers putts in an average round consist of taking two putting strokes to get the ball in the hole.

Strokes Gained Putting

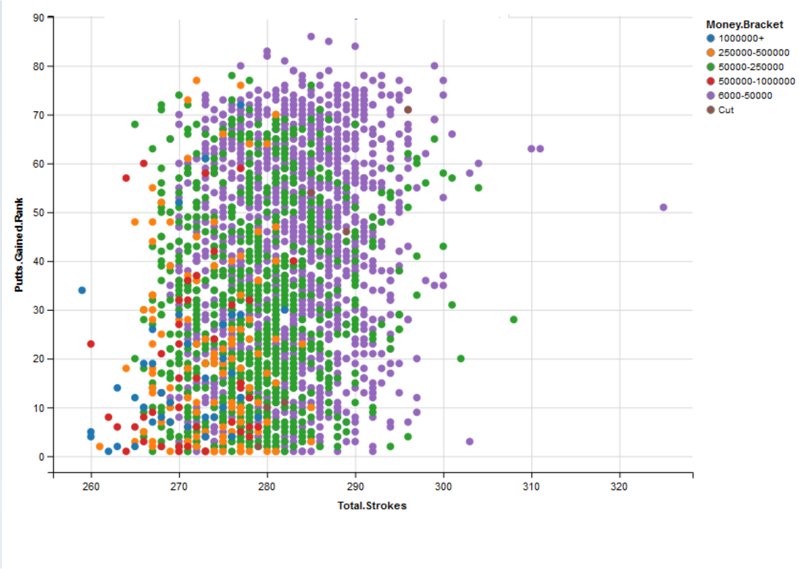


Figure 23 - Scatterplot of Birdies vs. Putts Gained clustered by Money Bracket

A new statistic created by the use of ShotLink data. It is reported to be the best measure for putting on Tour, measuring how many strokes a player gained against the field. The above figure 23 illustrates the importance of Strokes Gained Putting to Total Strokes and Money Earned. It can be seen that players who earned over $1,000,000 in one event were never ranked outside the top 40 for Strokes Gained Putting that week. The scatterplot also tells us that the higher the individual ranks for Strokes Gained Putting, the less money they are likely to earn.

6.6 Scrambling

For when an individual does not hit the Green in Regulation, if they achieve a score of Par or Better they have successfully "Scrambled" on that hole. Positive scrambling stats are usually associated with players who possess more finesse and accuracy with respect to those who base their game on power.

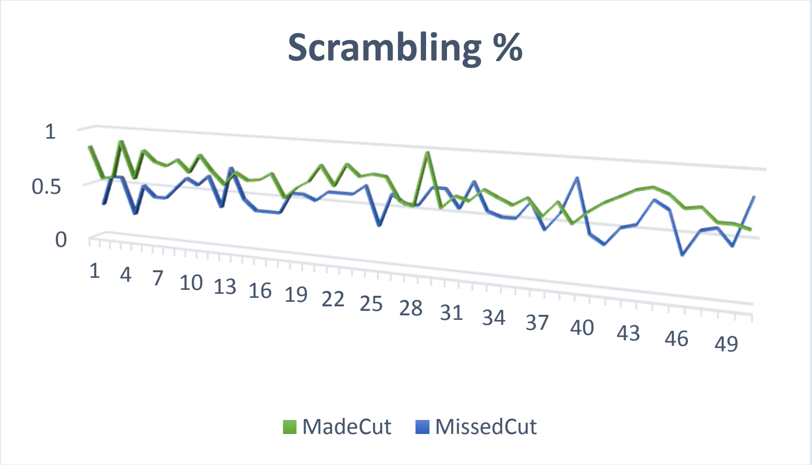


Figure 24 - Line Chart of Scrambling% of players who Missed Cut and Made Cut

In 2016 season there was clear evidence to suggest that Scrambling had an impact on player performance and success. On average players who made the Cut successfully scrambled 60.87% of the time, compared to 50.81% of players who missed the Cut. Figure 24 is based on a random sample of 50 observations (x-axis) of players who Made and Missed the Cut. The line chart emphasizes the influence of scrambling when attempting to make the Cut. In one instance, a player successfully scrambled 100%; this was Brian Stuard in the Zurich Classic who eventually finished 1st.

6.7 Tiger Woods Scandal

Anyone familiar with the world of sport will know that Tiger Woods was one of the most dominant sportsmen throughout the 2000's. In December 2009, news was released that Woods was involved in an extramarital affair. Let's test a theory if Woods performance changed before and after this news broke out. Taking performance data from 2007 to 2009 and 2010 to 2012 we will test if there was any difference in his performance before and after the scandal. Let's take the amount of Money earned as the explanatory variable for this test. This does not account for any sponsored related earnings, purely tournament winnings. A Paired T-Test will be used for this experiment.

Paired T-Test

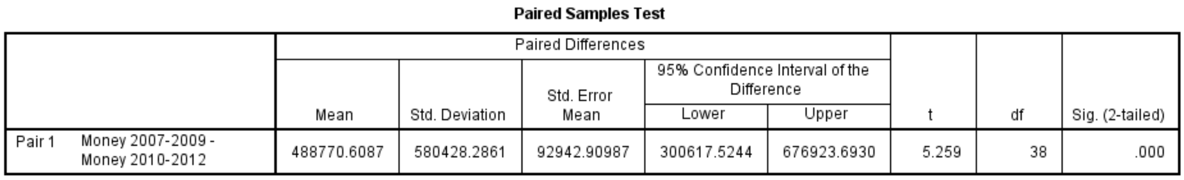
Null Hypothesis – H0: µ (Woods 07-09) = µ (Woods 10-12)

There is no difference in the mean of Money earned by Tiger Woods before and after the scandal that revolved around him.

Alternative Hypothesis – H1: µ (Woods 07-09) ≠ µ (Woods 10-12)

There is a difference in the mean of Money earned by Tiger Woods before and after the scandal that revolved around him.

Table 10 - Paired Samples Test Tiger Woods



Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	Money 2007-2009 - Money 2010-2012	488770.6087	580428.2861	92942.90987	300617.5244	676923.6930	5.259	38	.000

In table 10 we can see the Paired T-Test results. A P-Value of <0.001 was reported.

T = 5.259, df = 38, P-Value <0.001

We reject the Null Hypothesis in favour of the Alternative that there is a difference in the mean of Money earned by Tiger Woods before and after the scandal that revolved around him. If we examine the boxplot in figure 25 there is clear evidence to support the theory that Tiger Woods performance was in fact affected by the scandal that broke out. A significantly higher amount of money was earned in total between 2007 and 2009 ($27,150,214) compared to ($8,088,158) between 2010 and 2012, (Calculated in RStudio). There is also a noticeable difference between the two medians and 1^{st} and 3^{rd} quartile ranges. There were only two events where Woods earned over $1,000,000 following the scandal.

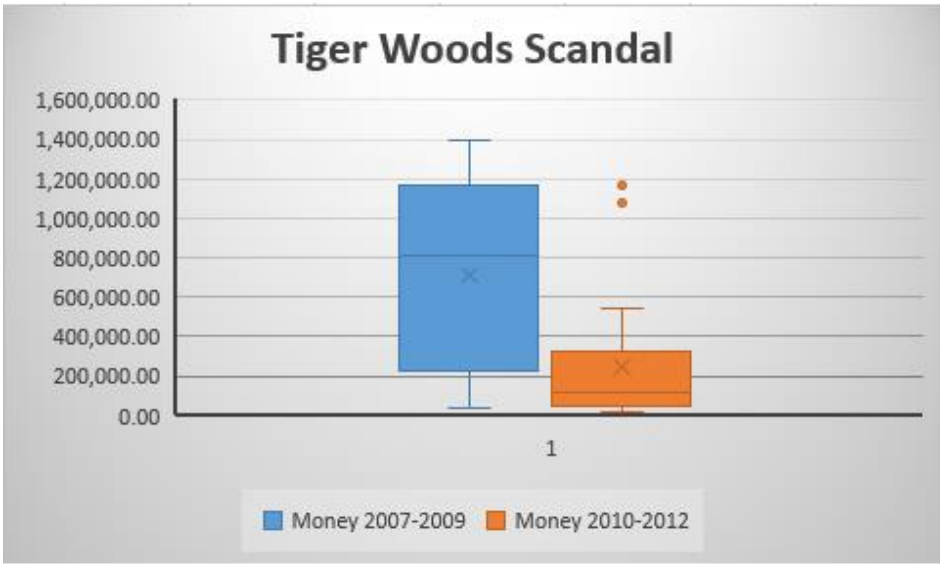


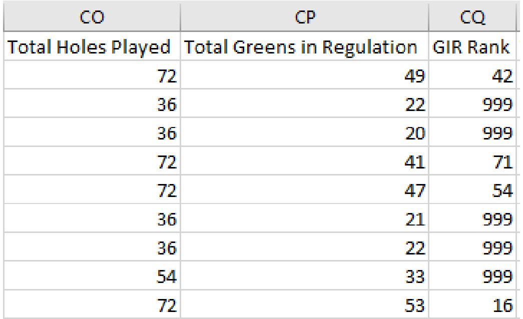
Figure 25 - Woods Earnings

7. Predicting Golfer Performance

Predicting Golfer performance is a difficult task due to the massive variety of Variables, Players, Observations and inconsistency within the data.

The most difficult problem when attempting to predict results for Golf Tournaments is the inconsistency with the data. Firstly, let's take an example to put it into perspective. If you wanted to predict how much an individual had to pay for car insurance or the prices of housing, you would take the variables dependant on (Insurance or Pricing) to use in your Model / Algorithm. The difference in the data is that there is less inconsistency within the examples given; previous pricing for houses / insurance would be more homogeneous compared to the Shotlink data which is essentially determined by whether a player makes the cut or not. This leads to previously mentioned issues such as the "999" value.

Table 11 - Data Sample



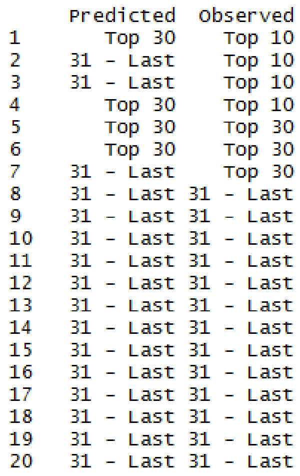
CO	CP	CQ
Total Holes Played	Total Greens in Regulation	GIR Rank
72	49	42
36	22	999
36	20	999
72	41	71
72	47	54
36	21	999
36	22	999
54	33	999
72	53	16

The sample of data in table 11 illustrates the inconsistency of the dataset. We can see that when a player makes the cut (72 Holes Played) they will have a rank based on how they performed in that category against the field for that week. When a player misses the cut they are assigned a value of "999" fr any field that accounts for "rank". When measuring the number of Total Greens in Regulation for the individual we can see that in most cases players who played 72 holes tend to have twice as many "Greens in Regulation" to those who played 36 holes. Dividing / Multiplying the values by two would not be sufficient in this case as it would not provide an inaccurate depiction of how a player who missed the cut would have hypothetically played 72 holes instead of 36. Instead I have calculated such relevant variables as a percentage to provide a fairer assessment of how a player performed that week regardless of how many holes they played. There are several instances where there was 54 holes played, due to weather affected tournaments or if a 3 round cut was implemented for that tournament. These observations were removed in some cases to allow for a more accurate model.

7.1 KNN (K Nearest Neighbours)

KNN was used in attempt to predict the finishing bracket of a sample of 1030 observations. The original data was split into a train and test set for the purposes of this model, 67% to 33% respectively. Typically, the k value would be selected by calculating the square root of training instances. In this case k was too large and did not prove sufficient when testing. After testing several k values, it was discovered that a k value of 7 was best suited to the model. The fact k usually falls between 3 and 10 was taken into consideration for this selection.

A sample of results of the KNN model are illustrated in figure 25. The results symbolize the NN classifiers of Finish Bracket for the test set against the predicted results. Figure 26 represents an evaluation of the model. At first look there is no evidence to suggest the model is highly accurate.

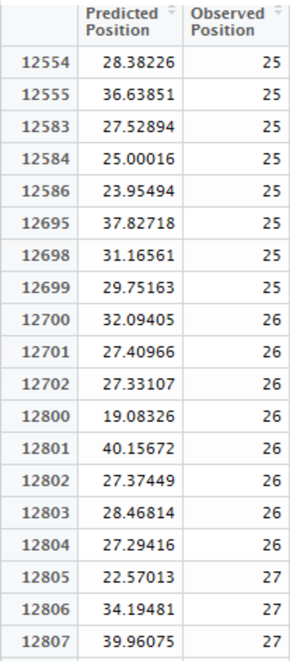


```
     Predicted  Observed
1       Top 30    Top 10
2    31 - Last    Top 10
3    31 - Last    Top 10
4       Top 30    Top 10
5       Top 30    Top 30
6       Top 30    Top 30
7    31 - Last    Top 30
8    31 - Last 31 - Last
9    31 - Last 31 - Last
10   31 - Last 31 - Last
11   31 - Last 31 - Last
12   31 - Last 31 - Last
13   31 - Last 31 - Last
14   31 - Last 31 - Last
15   31 - Last 31 - Last
16   31 - Last 31 - Last
17   31 - Last 31 - Last
18   31 - Last 31 - Last
19   31 - Last 31 - Last
20   31 - Last 31 - Last
```

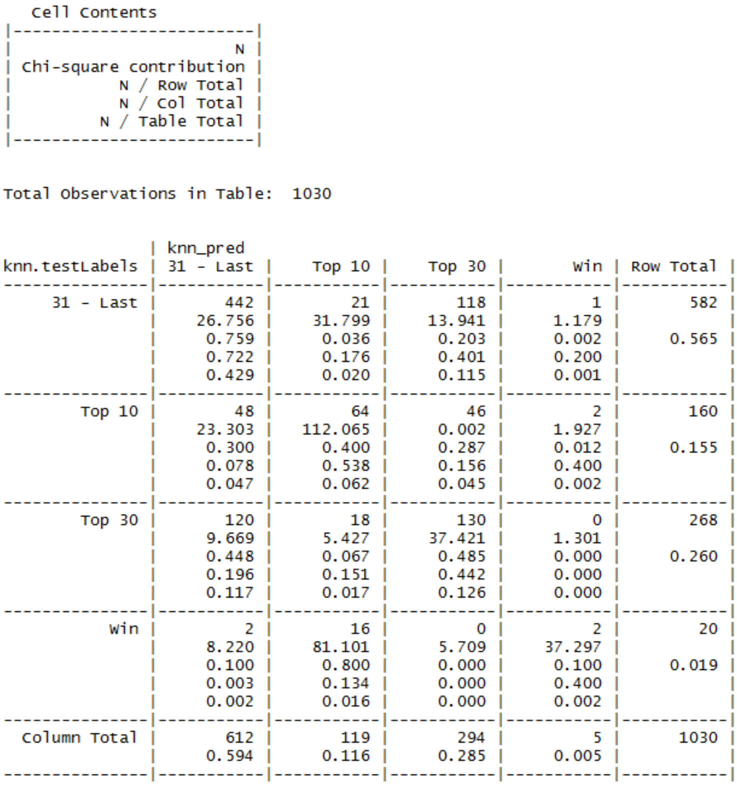
Figure 26 - Screenshot of KNN results in R

The model was most accurate when predicting if a golfer finished between 31st and last position, successfully predicting 442 out of 582 observations. 64 Top 10 finishes were successful predicted out of 164 observations. 130 of the 268 Top 30 observations were successfully predicted. 2 of the 20 Wins were predicted. As these results are based on data of events that already happened the model is not accurate enough to predict events that have not taken place. We can also test for a predicted Finish Position as seen in table 12.

Table 12-Valero KNN Results



	Predicted Position	Observed Position
12554	28.38226	25
12555	36.63851	25
12583	27.52894	25
12584	25.00016	25
12586	23.95494	25
12695	37.82718	25
12698	31.16561	25
12699	29.75163	25
12700	32.09405	26
12701	27.40966	26
12702	27.33107	26
12800	19.08326	26
12801	40.15672	26
12802	27.37449	26
12803	28.46814	26
12804	27.29416	26
12805	22.57013	27
12806	34.19481	27
12807	39.96075	27



```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  1030


               | knn_pred
knn.testLabels |  31 - Last |     Top 10 |     Top 30 |        Win | Row Total |
---------------|------------|------------|------------|------------|-----------|
     31 - Last |        442 |         21 |        118 |          1 |       582 |
               |     26.756 |     31.799 |     13.941 |      1.179 |           |
               |      0.759 |      0.036 |      0.203 |      0.002 |     0.565 |
               |      0.722 |      0.176 |      0.401 |      0.200 |           |
               |      0.429 |      0.020 |      0.115 |      0.001 |           |
---------------|------------|------------|------------|------------|-----------|
        Top 10 |         48 |         64 |         46 |          2 |       160 |
               |     23.303 |    112.065 |      0.002 |      1.927 |           |
               |      0.300 |      0.400 |      0.287 |      0.012 |     0.155 |
               |      0.078 |      0.538 |      0.156 |      0.400 |           |
               |      0.047 |      0.062 |      0.045 |      0.002 |           |
---------------|------------|------------|------------|------------|-----------|
        Top 30 |        120 |         18 |        130 |          0 |       268 |
               |      9.669 |      5.427 |     37.421 |      1.301 |           |
               |      0.448 |      0.067 |      0.485 |      0.000 |     0.260 |
               |      0.196 |      0.151 |      0.442 |      0.000 |           |
               |      0.117 |      0.017 |      0.126 |      0.000 |           |
---------------|------------|------------|------------|------------|-----------|
           Win |          2 |         16 |          0 |          2 |        20 |
               |      8.220 |     81.101 |      5.709 |     37.297 |           |
               |      0.100 |      0.800 |      0.000 |      0.100 |     0.019 |
               |      0.003 |      0.134 |      0.000 |      0.400 |           |
               |      0.002 |      0.016 |      0.000 |      0.002 |           |
---------------|------------|------------|------------|------------|-----------|
  Column Total |        612 |        119 |        294 |          5 |      1030 |
               |      0.594 |      0.116 |      0.285 |      0.005 |           |
---------------|------------|------------|------------|------------|-----------|
```

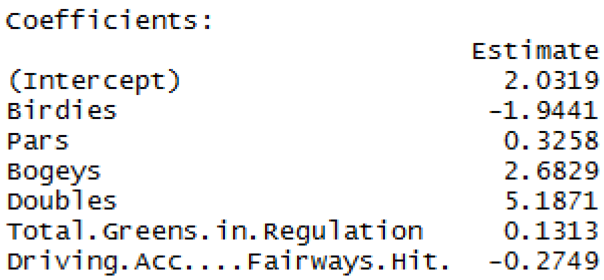
Figure 27 - KNN Results

When examining the results of predicting finish position, it is worth noting that the results are more consistent within the mid ranked results 30th – 40th. One explanation for this is that there is more consistency where this data lies as the average finish position is 36th on the PGA Tour.

7.2 Multiple Linear Regression

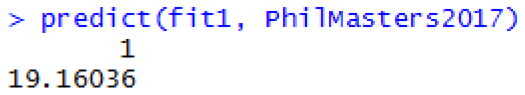
Masters 2017

Table 13 - Orignial Masters Model



```
Coefficients:
                              Estimate
(Intercept)                     2.0319
Birdies                        -1.9441
Pars                            0.3258
Bogeys                          2.6829
Doubles                         5.1871
Total.Greens.in.Regulation      0.1313
Driving.Acc....Fairways.Hit.   -0.2749
```

Table 13 is a Multiple Linear Regression Model based on Masters data from 2010 to 2016. Let's use the model to make a prediction for Phil Mickelson in this year's Masters. The co-efficient for the model can also be used by the players to enter predicted values for the independent variables which will predict a finish position (intercept value) for that event. First by calculating Phil Mickelson's average results for the independent variables we can gather an interpretation of how the individual has performed in this event in past circumstances. A new data frame must be created with these values. The data frame consists of only the independent variables with an empty value for the (Finish Position) to be predicted based on the model which measures data for 2010 – 2016. By using the predict function in R to predict the model against the Phil Mickelson 2017 data (event yet to happen) we return a result of:



```
> predict(fit1, PhilMasters2017)
       1
19.16036
```

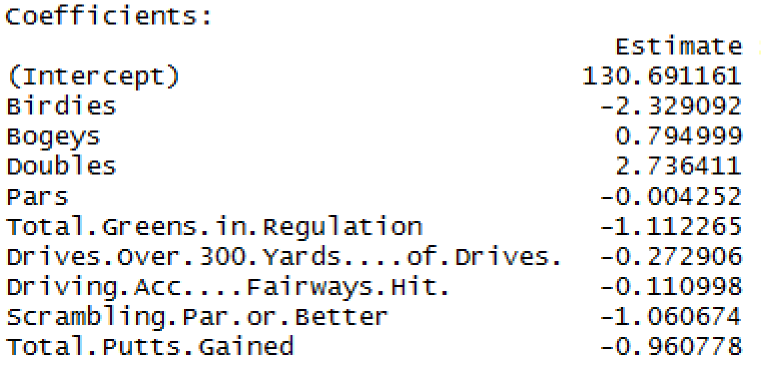
After the event took place, Phil Mickelson eventually finished 22nd. Quite an accurate prediction. With one example not been enough to evaluate the success of the model let's further examine if there is a way to gather a higher success rate with multiple observations.

Before we examine if the model can handle multiple observations let's add some independent variables to the model and test for a new tournament. This time the "Valero Texas Open" will be the tournament of choice due to the larger amount of historical data. For each separate tournament a new model must be created as each tournament's variables for success differ, as they are played on different courses with varied lengths, difficulties and layouts. The independent variables will stay the same every time, but the model will be based off different historical data (Valero Texas Open 2010-2016). The variables been selected for the model are those that are most influential to success of a professional golfer on the PGA Tour.

7.3 Tournament Predictions

7.3.1 Valero Texas Open

Table 14 - Valero Texas Open Model Co-Efficients



```
Coefficients:
                                         Estimate
(Intercept)                            130.691161
Birdies                                 -2.329092
Bogeys                                   0.794999
Doubles                                  2.736411
Pars                                    -0.004252
Total.Greens.in.Regulation              -1.112265
Drives.over.300.Yards....of.Drives.     -0.272906
Driving.Acc....Fairways.Hit.            -0.110998
Scrambling.Par.or.Better                -1.060674
Total.Putts.Gained                      -0.960778
```

By adding the above variables, we have increased the R – Squared value to 89.68% and an Adjusted R – Square Value of 89.55%. Some values have also been changed to provide a more accurate result. The Finish Position value of “999” when missing the cut has been altered to “100” to provide a more accurate model. I have selected 100, as the final finish position when the cut is made is 89th.

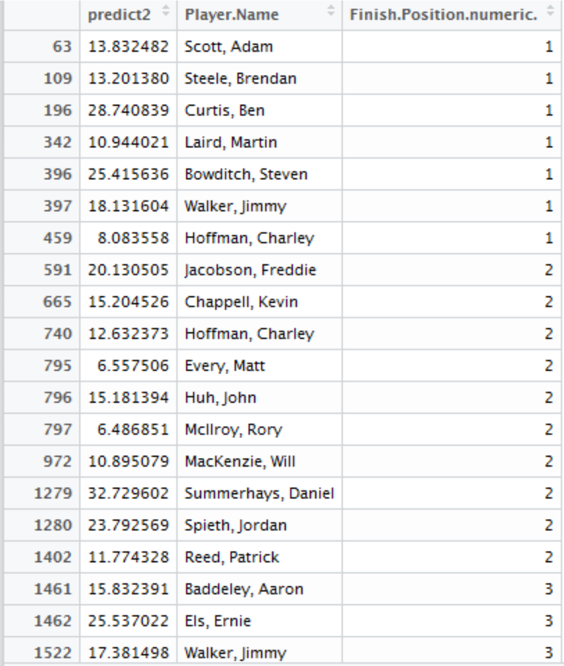
First let’s test the new model against the results of the “Valero Texas Open” from 2010 to 2016. From table 15 we can see that there is a somewhat of a correlation between the predicted results and the observed results (Finish.Position.numeric).

After testing for correlation we can see that there is in fact a very strong correlation between the Predicted results and the observed results for the Valero Texas Open 88.11%.

cor(predict2, Valero[1:730, 14])

[1] 0.8811194

Table 15 - Valero 2016 Observed Prediction Results



	predict2	Player.Name	Finish.Position.numeric.
63	13.832482	Scott, Adam	1
109	13.201380	Steele, Brendan	1
196	28.740839	Curtis, Ben	1
342	10.944021	Laird, Martin	1
396	25.415636	Bowditch, Steven	1
397	18.131604	Walker, Jimmy	1
459	8.083558	Hoffman, Charley	1
591	20.130505	Jacobson, Freddie	2
665	15.204526	Chappell, Kevin	2
740	12.632373	Hoffman, Charley	2
795	6.557506	Every, Matt	2
796	15.181394	Huh, John	2
797	6.486851	McIlroy, Rory	2
972	10.895079	MacKenzie, Will	2
1279	32.729602	Summerhays, Daniel	2
1280	23.792569	Spieth, Jordan	2
1402	11.774328	Reed, Patrick	2
1461	15.832391	Baddeley, Aaron	3
1462	25.537022	Els, Ernie	3
1522	17.381498	Walker, Jimmy	3

Now let's test the model on the Valero Texas Open for 2017. Brendan Steele is the example to be used for this test. Steele won this event in 2011. First we take the averages of his historical data for this event, then we apply these values to the model.

```
predict(fit2, SteeleValero)
   1
38.3257
```

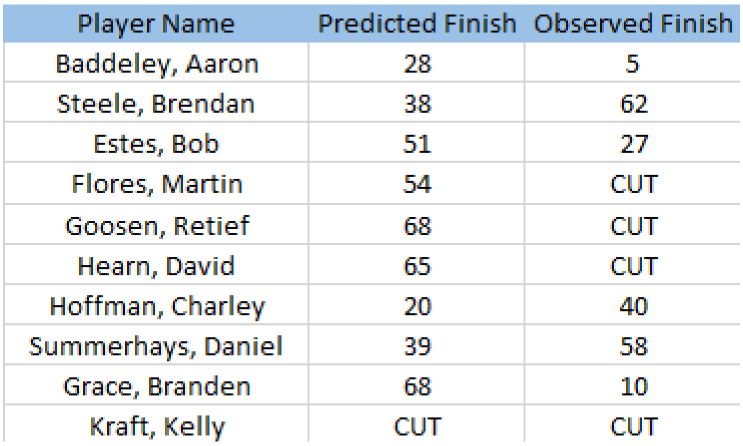
It is predicted that Brendan Steele will finish 38th during the 2017 Valero Texas Open. Steele eventually finished in a tie for 62nd.

Let's test another player who is due to play in the Valero Texas Open; Aaron Baddeley. It is predicted that Aaron Baddeley will finish 28th during the 2017 Valero Texas Open. He eventually finished 5th.

```
predict(fit2, AaronValero)
   1
28.23857
```

It is predicted that Aaron Baddeley will finish 28th during the 2017 Valero Texas Open. He eventually finished 5th. After testing for more instances in table 16 we can see that the predicted and observed results tend to vary. Any value of over 76 means the predicted player will miss the Cut. We can see that in the predicted instances of Retief Goosen and David Hearn they were close to missing the cut.

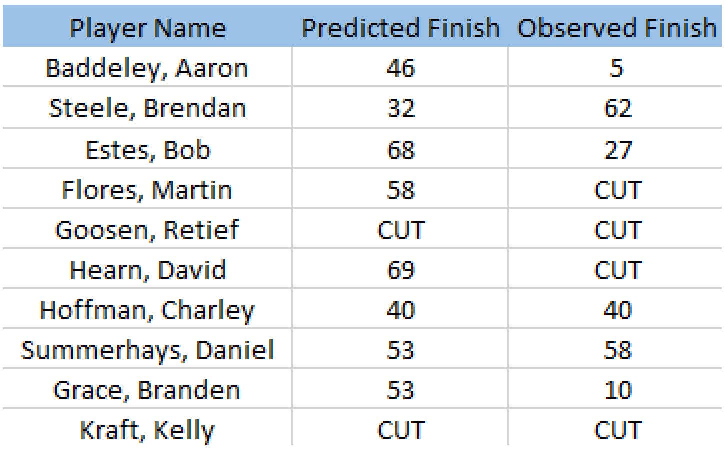
Table 16 - Orignal Valero Predictions



Player Name	Predicted Finish	Observed Finish
Baddeley, Aaron	28	5
Steele, Brendan	38	62
Estes, Bob	51	27
Flores, Martin	54	CUT
Goosen, Retief	68	CUT
Hearn, David	65	CUT
Hoffman, Charley	20	40
Summerhays, Daniel	39	58
Grace, Branden	68	10
Kraft, Kelly	CUT	CUT

As the model is based on only historical data for the Valero Texas Open, player form is not taken into account. Let's examine if there is a difference in the results if we implement how a player has performed in recent events into the model.

Table 17 - Valero 2017 Predictions



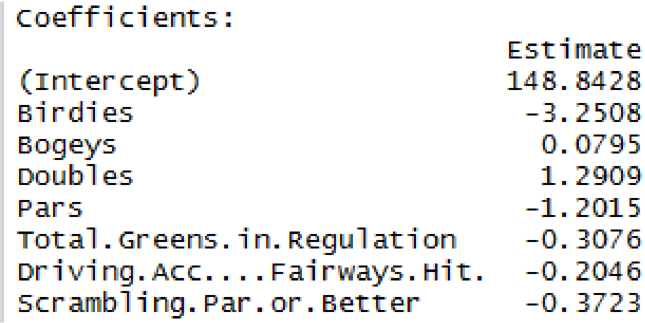
Player Name	Predicted Finish	Observed Finish
Baddeley, Aaron	46	5
Steele, Brendan	32	62
Estes, Bob	68	27
Flores, Martin	58	CUT
Goosen, Retief	CUT	CUT
Hearn, David	69	CUT
Hoffman, Charley	40	40
Summerhays, Daniel	53	58
Grace, Branden	53	10
Kraft, Kelly	CUT	CUT

By implementing the players recent form (Average Finish Position for previous few events) into the model, we can see that there is a difference in the results. Firstly, Charley Hoffman was predicted to finish 40th, a correct prediction. There was an accuracy increase in all predictions apart from the first 3 observations. An interpretation of Aaron Baddeley's results could be that his poor form in recent events was the reason he was predicted to finish higher than originally predicted. His eventual finish of 5th was more than likely a surprise. Aaron Baddeley had odds of 125/1 at the prior to the tournament, the favourite was 9/1. This was an indication of how his eventual finish was unsuspected.

7.3.2 Masters

It is well known that the Masters is one of the most prestigious events in Golfing History. If you ask any professional golfer, the majority will say that the Masters would be their dream tournament to win. A win at the Masters is accompanied by a Famous "Green Jacket". Let's update the model for the "Masters" and see what the results look like. Two Variables have been removed from the model "Drives Over 300 yards" and "Total Putts Gained". These variables were not recorded in ShotLink's data for previous Masters events. Table 18 provides the Multiple Linear Regression Model based on Masters data from 2010 – 2016.

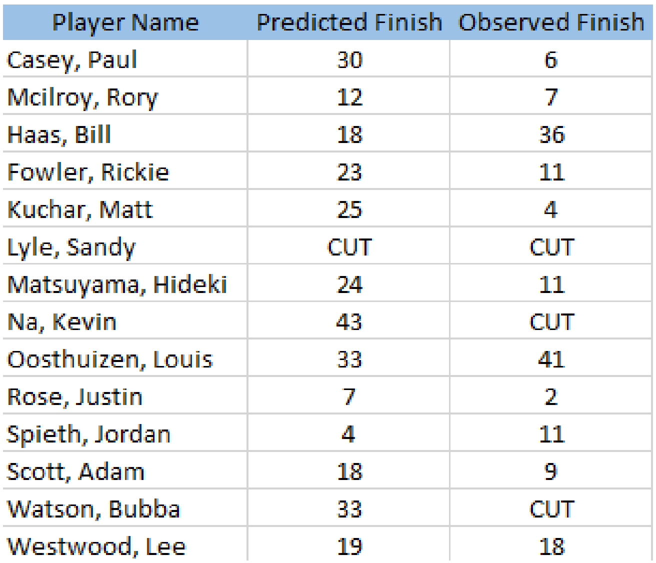
Table 18 - Model Co-efficient for Masters



```
Coefficients:
                               Estimate
(Intercept)                    148.8428
Birdies                         -3.2508
Bogeys                           0.0795
Doubles                          1.2909
Pars                            -1.2015
Total.Greens.in.Regulation      -0.3076
Driving.Acc....Fairways.Hit.    -0.2046
Scrambling.Par.or.Better        -0.3723
```

The model has a very high R – Squared Value of 94.55% and an adjusted R –Squared Value of 94.48%. In table 19 we can see there was a pretty high success in predicting the Results for the 2017 Masters. The observations were selected from players who have played the event at least twice between 2010 and 2016 and players who have played at least two events in the last 5 months. After this criterion was applied the observations were selected at random.

Table 19 - Masters Predictions 2017

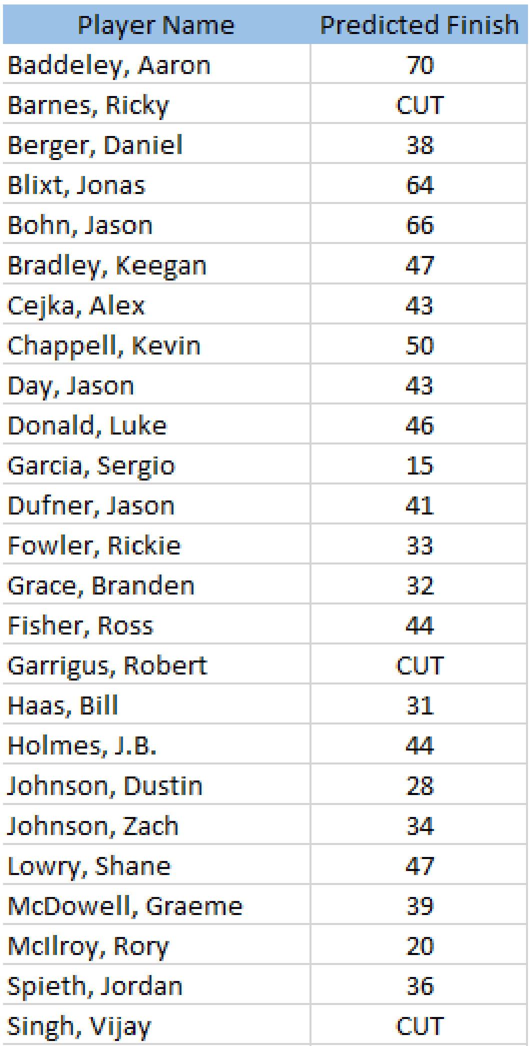


Player Name	Predicted Finish	Observed Finish
Casey, Paul	30	6
Mcilroy, Rory	12	7
Haas, Bill	18	36
Fowler, Rickie	23	11
Kuchar, Matt	25	4
Lyle, Sandy	CUT	CUT
Matsuyama, Hideki	24	11
Na, Kevin	43	CUT
Oosthuizen, Louis	33	41
Rose, Justin	7	2
Spieth, Jordan	4	11
Scott, Adam	18	9
Watson, Bubba	33	CUT
Westwood, Lee	19	18

7.3.3 Player’s Championship Predictions May 11th – 14th

The following predictions have been made for the Player’s Championship starting on the 11th May. The Player’s is considered the unofficial 5th Major and attracts a lot of interest from the top players in the world. Betting companies will also relish the opportunity of making money on a popular event. The predicted results are indicated in table 20.

Table 20 - Players Predictions 2017



Player Name	Predicted Finish
Baddeley, Aaron	70
Barnes, Ricky	CUT
Berger, Daniel	38
Blixt, Jonas	64
Bohn, Jason	66
Bradley, Keegan	47
Cejka, Alex	43
Chappell, Kevin	50
Day, Jason	43
Donald, Luke	46
Garcia, Sergio	15
Dufner, Jason	41
Fowler, Rickie	33
Grace, Branden	32
Fisher, Ross	44
Garrigus, Robert	CUT
Haas, Bill	31
Holmes, J.B.	44
Johnson, Dustin	28
Johnson, Zach	34
Lowry, Shane	47
McDowell, Graeme	39
McIlroy, Rory	20
Spieth, Jordan	36
Singh, Vijay	CUT

8. Conclusion

This section will conclude the findings of the project and if the stated objectives were achieved. The initial objectives will be summarised and assessed based on the results generated in this dissertation. A further evaluation of the predictive results will be discussed as well as some potential ideas for future research with ShotLink data or general Golf Statistics.

8.1 Overview of Research

As previously mentioned, ShotLink data has been used for different research purposes. Creation of new statistics has been the main focus for ShotLink Data. Attempting to predict and analyse success of a professional golfer on the PGA Tour using machine learning models and other statistical tools has not been done through the use of data provided by ShotLink.

This dissertation determines the key aspects between success and failure on the PGA Tour by measuring the critical components of golf score. The dissertation aims to answer proposed questions through hypothesis testing and statistical analysis. It has been completed with the intention to predict results for individuals in certain tournaments.

The performed research was used to help provide guidance and inspiration to the original idea. The research concentrated on data analytics in Golf and other sports as well as data analysis and mining as a whole. The research also helped to gain a further understanding of how ShotLink is represented and how the data is gathered and interpreted.

8.2 Evaluation of Analysis

After performing a detailed analysis of the ShotLink data at “event” level I have made several discoveries on the performance of players on the PGA Tour. The PGA Tour can be a lottery when it comes to results and earnings due to the spontaneity of golf as a sport. Money can be a key motivator for many golfers. In 2016, there were multiple circumstances where players earned over $1,000,000 for succeeding in just one event. Even at a minimum, players earn $6,000 for finishing last assuming they make the Cut. It is extremely common for players to lose their PGA membership for missing a number of Cuts and not making required earnings figure annually.

Age is one of the reasons why golf is such a beautiful game, anyone can play it. While there are high levels of athleticism among some of the top players, many professional golfers are just the average person whose trade happens to be Golf. The average age of a PGA Tour player is 33. Commonly, players who are between the mid 20’s and mid 30’s tend to have more birdies. The number of birdies a golfer makes tends to decline once they reach their mid-40’s; still young. Despite this, there is no difference in the mean finish position of players under 33 and players over 33.

How a player scores for four rounds determines how they will finish for that tournament. The averages between each round of scoring are very similar. Round 3 (Saturday) is the most influential day for scoring and provides evidence for the theory of “Moving Day”. Players playing well in Round 3 will typically finish higher than playing well in any other round where if a player has a bad round on Saturday it is tested to be the worst day to have a bad round on. Players who are leading after the first round will usually finish in the top 10 and will have a very high chance of making the cut.

From my analysis, I have interpreted what the key variables are in correspondence with scoring on the PGA Tour; Driving Distance and Accuracy, Putting, Scrambling, Birdies to Bogeys Ratio and Greens in Regulation. All of these critical components have been implemented into the machine learning models for predictions.

8.3 Evaluation of Predictions

Although an interpretation has been already made for the models, it was brief, so let’s examine the models in further detail and propose if any potential changes could be made for future purposes.

The predicted results generated from the models were quite interesting considering Golf is probably one of the hardest things to predict considering the average field size is 130 players, compared to a football match where there are 3 possibilities; Win, Lose, Draw. We can see from the Masters predicted results and eventual results that there was a significant resemblance. With a very strong Adjusted R – Squared value we can assume some level of accuracy. Results for the Valero Texas open were also quite significant with one observation predicted to the exact finish. While this is more than likely a co-incidence, there is still a relationship between the observed results and the predicted. If a player was to use this model the best interpretation they could take is to measure the predicted finish to a certain degree, for example if a player is predicted to finish 69th, in the case of David Hearn in the Valero Texas Open 2017 with a Cut line of 70 players, we can make an interpretation that the player will be on the border of making the Cut or missing it. This would be a more valid observation than implying that the player will finish in 69th position or near last place.

Of course there are several issues with the models. The previously mentioned “999” value for the cut can be difficult to interpret. The assigned value of 100 may have to be altered for tournaments with a smaller player field size and Cut mark. For the example of the Masters, 10 places were taken away from the model's predicted result to account for the smaller field in the Masters. It is also noticeable that many of the predicted values fall between certain values and that there are few players predicted to finish in top 10 positions. This is due to the difference in the data for the Cut, because 58% of data was for players who made the Cut and 42% of the data was for players who missed the cut; an uneven distribution. Another issue is when measuring player form the assigned cut value of "100". If a player's last two events were a win and a missed Cut by one shot, the form score might not provide an accurate depiction of how they are playing. Hypothetically if for say the "Masters" was the only golf event played every year or even every week, we could predict the results of player performance to a higher degree of accuracy based solely on their historical data. Due to the fact of different tournaments requiring different approaches and strategies. Weather is also an unpredictable factor because a player playing one week in bad weather could be more affected by form than a player who is playing another week in good weather. Psychological factors must also be taken into account. Another factor when measuring form is non ShotLink data cannot be interpreted into the model. I am referring to other golf Tour's such as the "European Tour". Regularly players of European descent such as "Rory Milroy" play on both the European and PGA Tour. Form on the European Tour cannot be calculated as ShotLink to not provide this data. It is also difficult to predict players who have no form or previous experience in such event, e.g. a player who is making their debut start. There was no multicollinearity when tested for, present in the models.

Aside from players using the model / results to their advantage, one could also use the results in an attempt to have some fun with some harmless bets in hope that the predicted results are correct for the predicted event.

8.4 Future Ideas

Due to time constrictions these changes cannot be made to the models but for future purposes the following changes could be potentially altered.

- Creating a more reliable measure for players who miss the Cut.
- Add a new factor to determine player form score.
- Create a more effective model that calculates the prediction in multiple groups rather than individually.
- Divide the data to an even 50/50 for players who made and missed the Cut.
- Try model the data with a different machine learning model, such as logistic regression, time series analysis or decision trees.

- See if there is a way to account for non-statistical factors such as how Tiger Woods' marriage scandal affected his performance.
- See if there is any correlation between betting odds and forecasted results.

8.5 Achieved Objectives

The following objectives were achieved in this dissertation:

- Review of ShotLink research history, data science in Golf and appliance in other Sports.
- Statistical Analysis of ShotLink data highlighting the key findings and answering frequently asked questions by the world of Golf.
- Design & Evaluation of Predictive Models.
- Predictions for results of PGA Tour events in 2017.

8.6 Evaluation of Software

RStudio, SPSS and Microsoft Excel were the main technologies used for this project. RStudio was extremely efficient for generating results and designing plots. SPPS was critical for experimenting with statistical tests and testing hypotheses. Microsoft Excel was very convenient for managing data and manually arranging / adding tables / variables.

The potential for these models will hopefully offer a glimpse for the appliance of data analysis in Golf for future use.

This draws a conclusion to the project I thoroughly hoped you enjoyed reading through it and found it interesting.

Bibliography

- Agresti, A., 2007. An Introduction to Categorical DataAnalysis. Wiley. Web. 10 May 2017.
- Alexander, D.L., Kern, W., 2005. Drive for Show and Putt for Dough? An Analysis of the Earnings of PGA Tour Golfers. J. Sports Econ. 6, 46–60. doi:10.1177/1527002503260797. Web. 10 May 2017.
- "Accenture 6 Nations Rugby - Home". *Accenture-rugby.com*. N.p., 2017. Web. 10 May 2017.
- Broadie, M., 2011a. Putts Gained - Measuring Putting on the PGA Tour. Web. 10 May 2017.
- Broadie, M., 2008. Assessing Golfer Performance Using Golfmetrics. Web. 10 May 2017.
- Determinants of Performance on the PGA Tour. (2008). 1st ed. Andrew Peters. Available at: http://org.elon.edu/ipe/f2%20andrew%20peters%20final.pdf [Accessed 10 May 2017].
- Fearing, D., Acimovic, J., Graves, S., 2010. How to Catch a Tiger: Understanding Putting Performance on the PGA Tour (SSRN Scholarly Paper No. ID 1538300). Social Science Research Network, Rochester, NY. [Accessed 10 May 2017].
- Forbes.com. (2017). *Forbes Welcome*. [online] Available at: https://www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/#4ab1f2114c1f [Accessed 10 May 2017].
- Han, J., Kamber, M., Pei, J., 2012. Data Mining Concepts and Techniques, Third. ed. Morgan Kaufmann. [Accessed 10 May 2017].
- Golf Analytics. (2017). *Is Distance or Accuracy Preferable Off the Tee?*. [online] Available at: https://golfanalytics.wordpress.com/2014/09/30/is-distance-or-accuracy-preferable-off-the-tee/ [Accessed 10 May 2017].
- Kercheval, A. (2017). *An Introduction to Machine Learning for Cybersecurity and Threat Hunting*. [online] Blog.sqrrl.com. Available at: http://blog.sqrrl.com/an-introduction-to-machine-learning-for-cybersecurity-and-threat-hunting [Accessed 10 May 2017].
- Lewis, M. (2011). *Moneyball*. 1st ed. New York: W.W. Norton.
- Predicting Professional Golfer Performance Using Proprietary PGA Tour "Shotlink" Data. (2015). 1st ed. Brian Leahy. Available at: http://arrow.dit.ie/cgi/viewcontent.cgi?article=1072&context=scschcomdis [Accessed 10 May 2017].
- PGATour. (2017). *Stats*. [online] Available at: http://www.pgatour.com/stats/ [Accessed 10 May 2017].

- rimunozc, V. (2017). *The KDD Process for Extracting Useful Knowledge from Volumes of Data | CEINE – Centro de Investigación en Inteligencia de Negocios*. [online] Ceine.cl. Available at: http://www.ceine.cl/the-kdd-process-for-extracting-useful-knowledge-from-volumes-of-data/ [Accessed 10 May 2017].
- Riccio Ph.D, L., 2012. THE BEST FAIRWAY BALL STRIKER ON TOUR ! [Accessed 10 May 2017].
- Schultz, J. (2017). *Predicting PGA Tour Scoring Average from Statistics Using Linear Regression*. [online] Big-Ish Data. Available at: https://bigishdata.com/2016/07/05/predicting-pga-tour-scoring-average-from-statistics-using-linear-regression/ [Accessed 10 May 2017].
- Scikit-learn.sourceforge.net. (2017). *Clustering — scikits.learn v0.6-git documentation*. [online] Available at: http://scikit-learn.sourceforge.net/0.5/modules/clustering.html [Accessed 10 May 2017].
- Sigmazone.com. (2017). *Roger Clemens, Barry Bonds, and Hypothesis Testing Case Study*. [online] Available at: http://www.sigmazone.com/Clemens_Bonds_HypothesisTest.htm [Accessed 10 May 2017].
- Shotlink.com. (2017). [online] Available at: http://www.shotlink.com/ [Accessed 10 May 2017].
- Variability of score and golf ball trajectory in elite golfers. (2013). 1st ed. [ebook] James Parker. Available at: https://www.diva-portal.org/smash/get/diva2:716452/FULLTEXT01.pdf [Accessed 10 May 2017].
- Verma, J. (2016). *Sports Research with Analytical Solution using SPSS*. 1st ed. John Wiley & Sons, p.231.
- Welcome to The Golf College That Will Teach You How to Be a Professional Tour Golfer. (2017). *How to Be a Professional Golfer: Why 98 Percent of Young Golfers Will Fail and What You Can Do to Prevent it*. [online] Available at: http://www.protourgolfcollege.com/news-blog/how-to-become-a-professional-golfer-why-98-percent-of-young-golfers-will-fail-and-what-you-can-do-to-prevent-it [Accessed 10 May 2017].

Appendix

Appendix A – Golf Overview

Forms of Scoring

In golf, there are several common scoring terms used for shooting a score on a hole with relation to par. For some reason the terms are named after birds, not too sure why but would be interested to know.

Birdie

A Birdie is when a golfer takes one stroke less than par to get the ball in the hole. For example, taking 2 strokes on a par 3.

Bogey

A Bogey is when someone takes one stroke more than par to get the ball into the hole. An example would be taking 6 strokes to get the ball in the hole with a value of par 5.

Eagle

An Eagle is when a golfer shoots a score two shots under the par. For example, a 2 on a par 4. This is a lot rarer than a Birdie, Par or Bogey.

Double Bogey

Perhaps not the most of creative of names, a double Bogey is simply twice that of a Bogey, when someone completes a hole in 2 shots over par. This would be more common than an Eagle.

Other forms of Scoring

Hole in One

When a player strikes the ball in the hole from the tee box, usually a par 3. Also known as an Eagle.

Albatross (Double Eagle)

When a player completes a hole in three strokes under the par. This is an extremely rare bird hence why they rarely occur to professional golfers. Generally, they tend to happen when a player scores a 2 on a par 5. In some rare circumstances, Hole in One's have been recorded on short par 4's.

Triple Bogey, Quadruple Bogey

Player completes a hole in 3 or 4 strokes over par, respectively.

Condor

Although it is a term that will not be relevant to this dissertation, it is an interesting term that not many golfers are familiar with. It is when a golfer would shoot a score four under par to the hole. This is almost impossible as it would take a hole in one on a par 5, which even for the best golfers in the world is a near 0% chance.

Formats of Golf

While there are several different ways to play golf these days including team events such as Ryder Cup which includes Fourballs and Foursomes I will only be discussing the formats used on the PGA Tour.

Stroke Play

Stroke Play is the main format used on the PGA Tour. It consists of each player playing a number of rounds against a field of other golfers on the same course. The player who completes all the rounds in the lowest strokes is the winner.

Match Play

Not as common as stroke play. Consists of two players playing each other and no one else. Whichever player has the lowest score wins the hole. If they shoot the same score on that given hole the hole is halved / squared. Whichever player wins the most holes wins the match. The tournament consists of several knockout rounds. The winner is the last man standing

Common Terms

Handicap

A system used in amateur golf that implies how good they are at golf. It allows a fair playing field among players of all levels. If a golfer plays off a handicap of 18 they would essentially deduct 18 shots off of their final gross score of a round, this would produce their net score. An amateur golfers handicap would decrease the better they get and vice versa. Typically, a golfer would look to turn professional once they reach below a handicap of zero. After that it begins to go +1, +2 etc.

The Cut Line

The cut is a score mark halfway through the tournament that allows certain players to continue on to play the rest of the tournament, the players who fall into the cut mark are no longer entitled to play any further part in said tournament also meaning they will fail to earn any prize money that week. Usually the top 70/65 players make it to the weekend (Final 2 Rounds), anyone not in this bracket does not.

Greens in Regulation

A common statistic used by the PGA Tour. If a player reaches the green and has a chance at a birdie putt or if they have reached the green in two strokes less than par they have played the "Green in Regulation". Otherwise if the player takes 1 shot to reach the green on a par 3 or two shots on a par 4.

Driving Accuracy

How many percentage of fairways a player hits off the tee, not including par 3's.

Fairway & Rough

The fairway is where a golfer ideally wants to be, it is the shortest cut grass and even. The rough is usually thicker and longer making it more difficult to play from.

Strokes Gained Putting

A relatively new statistic brought in by PGA Tour through the use of research with ShotLink data. This statistic measures how many strokes a player gained against the field.

Tee to Green

Measures how well the player performed hitting the ball from the tee to the green, excludes putting.

Scrambling

If a golfer misses the green in regulation and score a par they have successfully scrambled. Measures the percentage of pars made when greens are missed.

The PGA Tour

The PGA Tour organises all professional golfing events played in the United States and North America. While the PGA offers several tiers of professional golf Tour's, the PGA Tour is the flagship series. It consists of around 35 – 40 events per year.

FedEx Cup

The FedEx cup is a season ending set of tournaments where the best players on the PGA Tour that season compete to win a prize of $10 million.

Majors

Majors are the four most prestigious golfing tournaments that take place every year, The U.S Open, Masters, PGA Championship and The Open. Winning a major is the pinnacle of any players career.

Appendix B – Project Proposal

Initial Project Proposal

Objectives

Background

Technical Approach

Special Resources Required

Project Plan

Evaluation

Project Objectives

The primary objective of this project is to analyse and predict trends on the success of professional golfers on the PGA (Professional Golfers Association) Tour through a detailed analysis of ShotLink ® Database. The data will be approached with statistical analysis methods of my choosing. The overall goal will be to project a result that highlights how professional golfers can improve their success on the PGA Tour. There are several steps required to complete this project successfully. In order to do this, I must begin with the project proposal. This proposal will outline my initial plans and ideas regarded for the project. While these plans and ideas may vary overtime, it will be a general guideline to base the project off for the coming future. I will also have to complete other activities in aid of completion of my project such as a requirements specification, conduct project analysis, perform a mid-point presentation, submit final documentation, perform a final presentation and display the project at a showcase. All these activities will be highlighted in more detailed project plan later.

The objective of this project as highlighted before is to predict trends and patterns in the success of the performance of professional golfers on the PGA Tour. I will be scrutinising a collection of data from the PGA Tour based on the last 15 years. I will be formulating several theories and testing several hypotheses. The process of this analysis will feature the use of ShotLink ® Database. I chose this database after an abundance of research. ShotLink is a platform that is used collectively with the PGA Tour to collect statistical and scoring data based on every shot of every player in real time making it a revolutionary platform. During my research, I came across a collection of data resources I could use to report my findings on but ultimately, I settled on ShotLink as it had a substantial selection of vastly detailed data. One obstacle that had to become overcome was gaining access to the database. In 2005 ShotLink allowed people to gain access to their database for academic purposes. After my submission request was approved by ShotLink and the PGA Tour I was granted access to the database set with rules and regulations needed to be accepted on my behalf.

As this project progresses I will be keeping a specific timetable in order to maintain a consistent work progress and avoid any unwanted circumstances such as late deliverable submissions or poor work quality. If I keep to a project plan it will aid me in completing a successful project.

Background

When I first realised that I had a choice to compose a final year project in a stream of my choice my decision was immediately to select the Data Analytics stream as the idea of measuring large quantities of data and working with statistics has always interested me. Once I selected this stream my next goal was to decide what to focus my project on. Been a major golf enthusiast I decided to focus on this area. The performance of a professional golfer is majorly based on statistics. I concluded to focus my project idea on attempting to predict trends and patterns in the success of the performance of a professional golfer on the PGA Tour. The influence for my decision came from both a personal interest point of view and to see if I can attempt to prove any theories or ideas on why certain statistics can lead to certain golfers being successful on the PGA Tour. For every individual attempting a project in the college there is a vetting process for the projects where each student would partake in a short presentation on why they should be allowed to attempt their project idea. There were 3 judges, if you got majority decision your idea was accepted if not your idea was rejected and you would have to focus on an idea suggested by the college lecturers.

Technical Approach & Details

Once I had finalised my project idea the next step was to decide how I would approach the task from a technical perspective. This involved heaps of research into what would be the best way to analyse the project idea from a technical perspective. The first approach I took was where I was going to obtain my data from which I have previously explained (ShotLink.) Now that I have been granted access to ShotLink the next step would be to perform an extremely detailed analysis review of the data I have access to. Once I have analysed this data thoroughly I will then select a multitude of variables on which I believe to be appropriate for testing and measuring the data at hand. Once all this data has been obtained and simplified to my liking I can then begin to test possible theories, trends and patterns within this data. Some ideas I currently have are to: create a new statistic for the PGA Tour to measure the performance of golfers on, for example currently there is a statistic for driving distance (how far a player hits the ball off the tee) for each individual player on the PGA Tour. My idea is to create an algorithm / formula which returns a result of a new form of statistic and propose it to the PGA Tour. I also plan to test several hypotheses such as comparing the results on the PGA Tour and contrasting that of betting odds or do world rankings and money earned justify these statistics. To analyse this data I plan to use several computer languages and software such as; SQL, R, Python, Excel, SPSS.

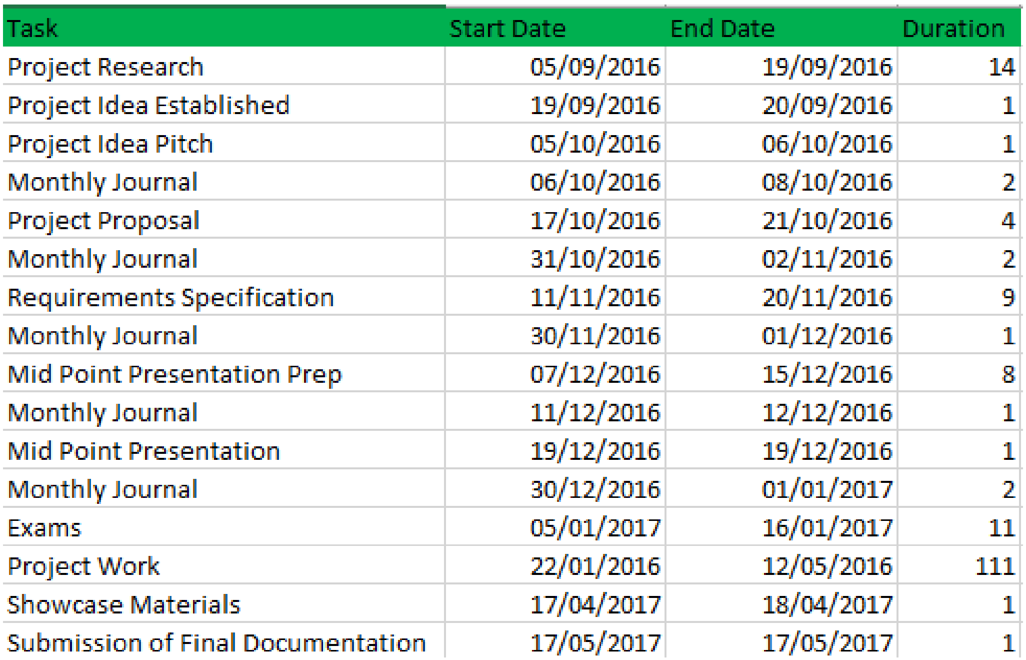
Special Resources Required

At the moment, I am currently lacking in knowledge of the computing languages R and Python. I plan to enhance my skills in this apartment by partaking in a module called Programming for Big Data. This module will teach all the key aspects of these languages and how to use them in relation to my project. If needs be I will also require several sources on the internet such as YouTube are programming blogs that will allow me to complete tutorials relevant to my understanding.

I will also need some basic resources to aid me in deliverance of this project:

- A basic PC with access to my student drive and college log in.
- A GitHub / Dropbox account to maintain any files I have.
- A USB to transfer and store files.
- Links to websites and blogs relevant to my project choice.

Project Plan



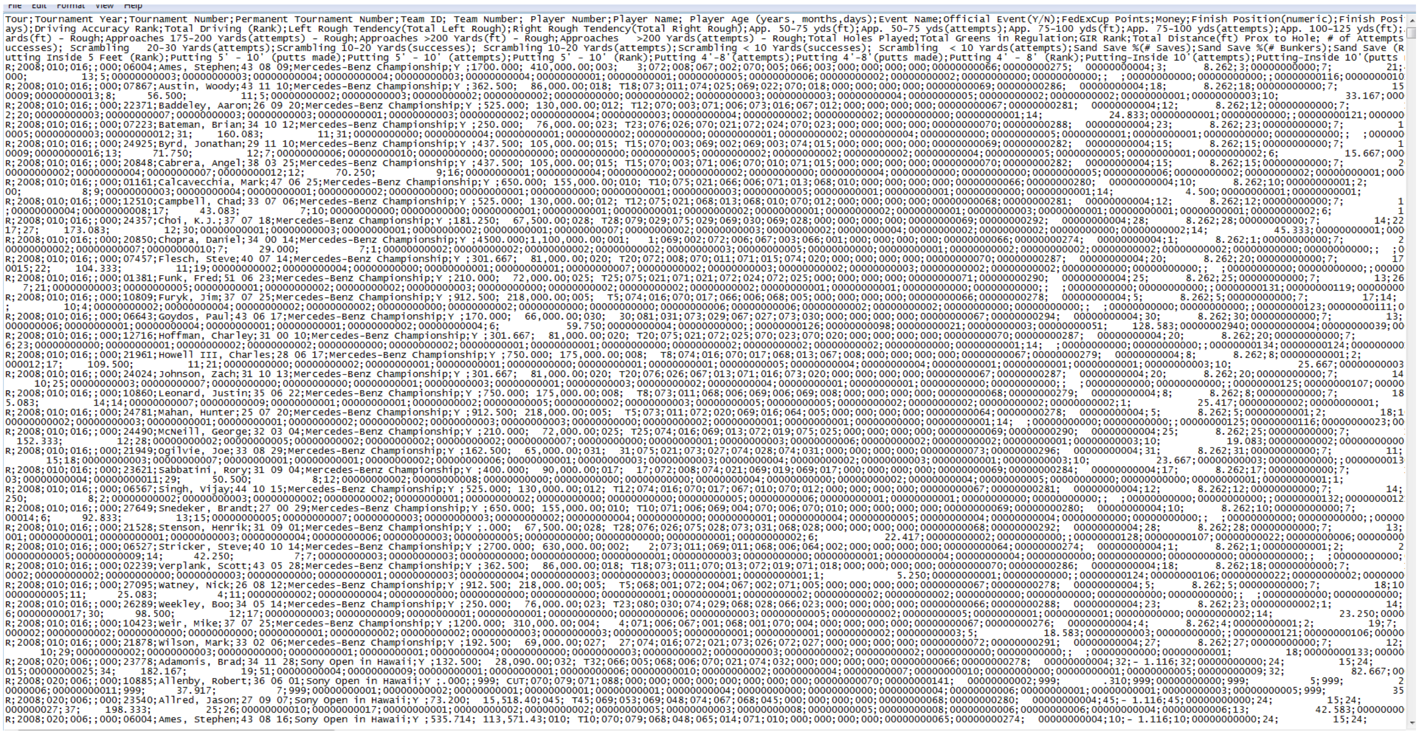
Task	Start Date	End Date	Duration
Project Research	05/09/2016	19/09/2016	14
Project Idea Established	19/09/2016	20/09/2016	1
Project Idea Pitch	05/10/2016	06/10/2016	1
Monthly Journal	06/10/2016	08/10/2016	2
Project Proposal	17/10/2016	21/10/2016	4
Monthly Journal	31/10/2016	02/11/2016	2
Requirements Specification	11/11/2016	20/11/2016	9
Monthly Journal	30/11/2016	01/12/2016	1
Mid Point Presentation Prep	07/12/2016	15/12/2016	8
Monthly Journal	11/12/2016	12/12/2016	1
Mid Point Presentation	19/12/2016	19/12/2016	1
Monthly Journal	30/12/2016	01/01/2017	2
Exams	05/01/2017	16/01/2017	11
Project Work	22/01/2016	12/05/2016	111
Showcase Materials	17/04/2017	18/04/2017	1
Submission of Final Documentation	17/05/2017	17/05/2017	1

Evaluation

This project plan will hopefully be used as a consistent guideline when engaging with any relevant project work. Though every mentioned beforehand in the project proposal may not necessarily be achieved I aim to maintain the plan provided as well as any ideas or suggestions I have proposed. I also aim to maintain the project plan I have set to even though I fully understand that are possibilities that I may come across stumbling blocks along the way which may or not interfere with the proposed deadlines.

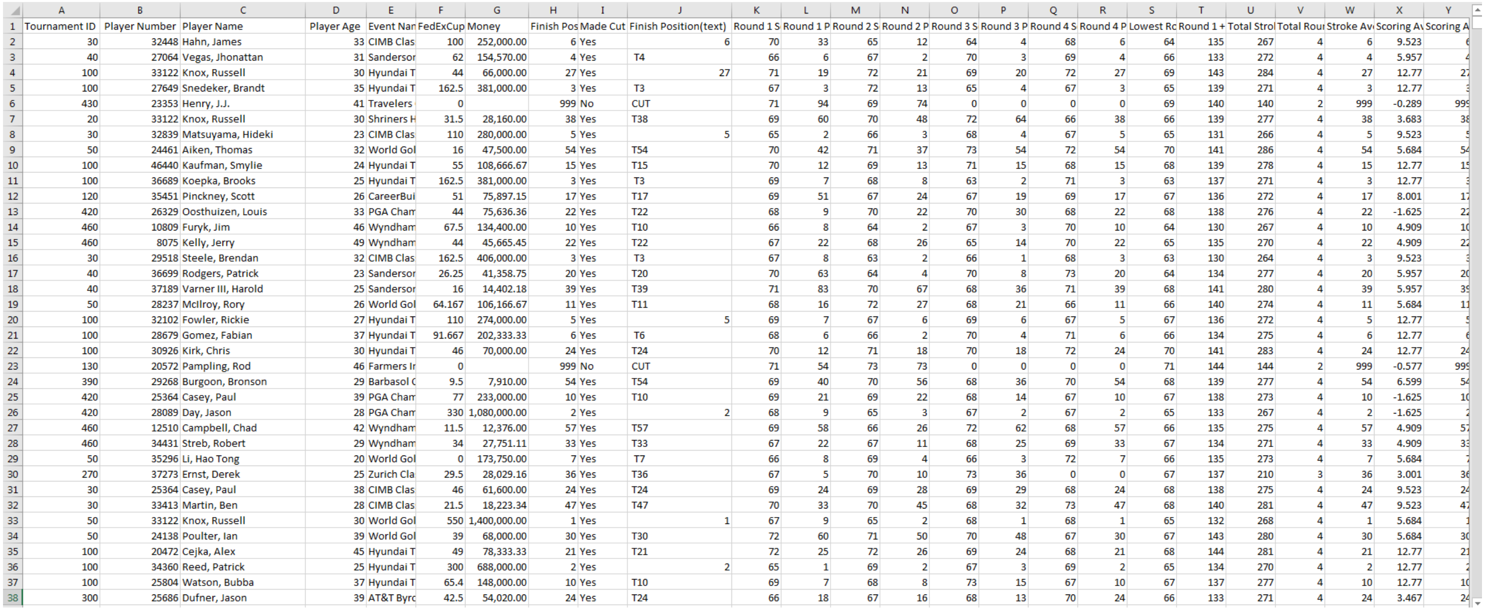
Appendix C - Data

Screenshot of ShotLink Data prior to Delimiter (.txt file)



```
Tour;Tournament Year;Tournament Number;Permanent Tournament Number;Team ID; Team Number; Player Number;Player Name; Player Age (years, months,days);Event Name;Official Event(Y/N);FedExCup Points;Money;Finish Position(numeric);Finish Posi
ays);Driving Accuracy Rank;Total Driving (Rank);Left Rough Tendency(Total Left Rough);Right Rough Tendency(Total Right Rough);App. 50-75 yds(ft);App. 50-75 yds(attempts);App. 75-100 yds(ft);App. 75-100 yds(attempts);App. 100-125 yds(ft);
ards(ft) - Rough;Approaches 175-200 Yards(attempts) - Rough;Approaches >200 Yards(ft) - Rough;Approaches   >200 Yards(attempts) - Rough;Total Holes Played;Total Greens in Regulation;GIR Rank;Total Distance(ft) Prox to Hole; # of Attempts
uccesses); Scrambling   20-30 Yards(attempts);Scrambling 10-20 Yards(successes); Scrambling 10-20 Yards(attempts);Scrambling < 10 Yards(successes); Scrambling  < 10 Yards(attempts);Sand Save %(# Saves);Sand Save %(# Bunkers);Sand Save (R
utting Inside 5 Feet (Rank);Putting 5' - 10' (putts made);Putting 5' - 10' (attempts);Putting 5' - 10' (Rank);Putting 4'-8'(attempts);Putting 4'-8'(putts made);Putting 4' - 8' (Rank);Putting-Inside 10'(attempts);Putting-Inside 10'(putts
R;2008;010;016;;000;06004;Ames, Stephen;43 08 09;Mercedes-Benz Championship;Y ;1700.000; 410,000.00;003;    3;072;008;067;002;070;005;066;003;000;000;000;000;00000000066;00000000275;   00000000004;3;        8.262;3;00000000000;7;           21;
000;          13;5;00000000003;00000000003;00000000004;00000000004;00000000003;00000000004;00000000001;00000000001;00000000005;00000000006;00000000002;00000000002;00000000000;00000000000;;   ;00000000000;00000000000;;00000000116;0000000010
R;2008;010;016;;000;07867;Austin, Woody;43 11 10;Mercedes-Benz Championship;Y ;362.500;  86,000.00;018; T18;073;011;074;025;069;022;070;018;000;000;000;000;00000000069;00000000286;   00000000004;18;       8.262;18;00000000000;7;           15
0009;00000000013;8;     56.500;        11;5;00000000002;00000000003;00000000002;00000000002;00000000000;00000000002;00000000003;00000000003;00000000004;00000000005;00000000002;00000000002;00000000001;00000000003;10;           33.167;000
R;2008;010;016;;000;22371;Baddeley, Aaron;26 09 20;Mercedes-Benz Championship;Y ;525.000; 130,000.00;012; T12;070;003;071;006;073;016;067;012;000;000;000;000;00000000067;00000000281;   00000000004;12;       8.262;12;00000000000;7;
2;20;00000000003;00000000007;00000000003;00000000003;00000000001;00000000003;00000000002;00000000004;00000000003;00000000004;00000000002;00000000002;00000000000;00000000001;14;          24.833;00000000001;00000000000;;00000000121;000000
R;2008;010;016;;000;07223;Bateman, Brian;34 10 12;Mercedes-Benz Championship;Y ;250.000;  76,000.00;023; T23;076;026;070;021;072;024;070;023;000;000;000;000;00000000070;00000000288;   00000000004;23;       8.262;23;00000000000;7;           1
0005;00000000003;00000000012;31;    160.083;         11;31;00000000000;00000000004;00000000001;00000000002;00000000000;00000000001;00000000002;00000000004;00000000000;00000000005;00000000001;00000000001;00000000000;00000000000;;   ;000000
R;2008;010;016;;000;24925;Byrd, Jonathan;29 11 10;Mercedes-Benz Championship;Y ;437.500; 105,000.00;015; T15;070;003;069;002;069;003;074;015;000;000;000;000;00000000069;00000000282;   00000000004;15;       8.262;15;00000000000;7;           1
0009;00000000016;13;    71.750;        12;7;00000000006;00000000010;00000000000;00000000000;00000000000;00000000005;00000000002;00000000002;00000000002;00000000004;00000000005;00000000005;00000000001;00000000002;6;           15.667;000
R;2008;010;016;;000;20848;Cabrera, Angel;38 03 25;Mercedes-Benz Championship;Y ;437.500; 105,000.00;015; T15;070;003;071;006;070;010;071;015;000;000;000;000;00000000070;00000000282;   00000000004;15;       8.262;15;00000000000;7;           2
00000000002;00000000004;00000000007;00000000012;12;     70.250;          9;16;00000000001;00000000004;00000000002;00000000002;00000000000;00000000004;00000000000;00000000000;00000000005;00000000006;00000000002;00000000002;00000000001;000
R;2008;010;016;;000;01161;Calcavecchia, Mark;47 06 25;Mercedes-Benz Championship;Y ;650.000; 155,000.00;010; T10;075;021;066;006;071;013;068;010;000;000;000;000;00000000066;00000000280;   00000000004;10;       8.262;10;00000000001;2;
00;          8;9;00000000003;00000000004;00000000001;00000000002;00000000000;00000000001;00000000000;00000000001;00000000003;00000000005;00000000001;00000000001;00000000000;00000000001;14;             4.500;00000000001;00000000001;
R;2008;010;016;;000;12510;Campbell, Chad;33 07 06;Mercedes-Benz Championship;Y ;525.000; 130,000.00;012; T12;075;021;068;013;068;010;070;012;000;000;000;000;00000000068;00000000281;   00000000004;12;       8.262;12;00000000000;7;           1
;00000000004;00000000008;17;     43.083;          7;10;00000000000;00000000000;00000000001;00000000001;00000000001;00000000002;00000000001;00000000002;00000000001;00000000003;00000000001;00000000001;00000000001;00000000002;6;           1
R;2008;010;016;;000;24357;Choi, K.J.;37 07 18;Mercedes-Benz Championship;Y ;181.250;  67,500.00;028; T28;079;029;075;029;069;030;069;028;000;000;000;000;00000000069;00000000292;   00000000004;28;       8.262;28;00000000000;7;         14;22
17;27;    173.083;        12;30;00000000001;00000000003;00000000001;00000000002;00000000001;00000000007;00000000002;00000000003;00000000002;00000000004;00000000002;00000000002;00000000000;00000000002;14;          45.333;00000000001;000
R;2008;010;016;;000;20850;Chopra, Daniel;34 00 14;Mercedes-Benz Championship;Y ;4500.000;1,100,000.00;001;    1;069;002;072;006;067;003;066;001;000;000;000;000;00000000066;00000000274;   00000000004;1;       8.262;1;00000000000;7;           2
00000000002;00000000007;00000000010;7;     29.000;          7;1;00000000002;00000000002;00000000002;00000000002;00000000003;00000000005;00000000000;00000000001;00000000002;00000000002;00000000002;00000000002;00000000000;00000000000;;   ;0
R;2008;010;016;;000;07457;Flesch, Steve;40 07 14;Mercedes-Benz Championship;Y ;301.667;  81,000.00;020; T20;072;008;070;011;071;015;074;020;000;000;000;000;00000000070;00000000287;   00000000004;20;       8.262;20;00000000000;7;          17
0015;22;    104.333;        11;19;00000000002;00000000004;00000000000;00000000001;00000000001;00000000007;00000000002;00000000003;00000000002;00000000003;00000000002;00000000002;00000000000;00000000000;;   ;00000000000;00000000000;;00000
R;2008;010;016;;000;01381;Funk, Fred;51 06 23;Mercedes-Benz Championship;Y ;210.000;  72,000.00;025; T25;075;021;071;021;072;024;072;025;000;000;000;000;00000000071;00000000290;   00000000004;25;       8.262;25;00000000000;7;         13;26
   7;21;00000000003;00000000005;00000000001;00000000002;00000000002;00000000003;00000000000;00000000002;00000000002;00000000002;00000000001;00000000001;00000000000;00000000000;;   ;00000000000;00000000000;;00000000131;00000000119;0000000
R;2008;010;016;;000;10809;Furyk, Jim;37 07 25;Mercedes-Benz Championship;Y ;912.500; 218,000.00;005;  T5;074;016;070;017;066;006;068;005;000;000;000;000;00000000066;00000000278;   00000000004;5;       8.262;5;00000000000;7;          17;14;
;          10;4;00000000002;00000000004;00000000002;00000000002;00000000000;00000000002;00000000000;00000000000;00000000006;00000000006;00000000002;00000000002;00000000000;00000000000;;   ;00000000000;00000000000;;00000000123;00000000111;0
R;2008;010;016;;000;06643;Goydos, Paul;43 06 17;Mercedes-Benz Championship;Y ;170.000;  66,000.00;030;  30;081;031;073;029;067;027;073;030;000;000;000;000;00000000067;00000000294;   00000000004;30;       8.262;30;00000000000;7;          13;
000000006;00000000001;00000000004;00000000001;00000000001;00000000002;00000000004;6;          59.750;00000000004;00000000000;;00000000126;00000000098;00000000021;00000000003;00000000051;     128.583;00000002940;00000000004;00000000039;00
R;2008;010;016;;000;12716;Hoffman, Charley;31 00 10;Mercedes-Benz Championship;Y ;301.667;  81,000.00;020; T20;075;021;072;025;070;023;070;020;000;000;000;000;00000000070;00000000287;   00000000004;20;       8.262;20;00000000000;7;
6;23;00000000000;00000000001;00000000002;00000000002;00000000000;00000000002;00000000001;00000000001;00000000000;00000000002;00000000002;00000000002;00000000002;00000000000;00000000001;14;   ;00000000000;00000000000;;00000000134;00000000124;000000000
R;2008;010;016;;000;21961;Howell III, Charles;28 06 17;Mercedes-Benz Championship;Y ;750.000; 175,000.00;008;  T8;074;016;070;017;068;013;067;008;000;000;000;000;00000000067;00000000279;   00000000004;8;       8.262;8;00000000001;2;
000012;17;    109.500;        11;21;00000000000;00000000002;00000000001;00000000001;00000000000;00000000001;00000000001;00000000005;00000000004;00000000004;00000000001;00000000001;00000000001;00000000003;10;          25.667;00000000003
R;2008;010;016;;000;24024;Johnson, Zach;31 10 13;Mercedes-Benz Championship;Y ;301.667;  81,000.00;020; T20;076;026;067;013;071;016;073;020;000;000;000;000;00000000067;00000000287;   00000000004;20;       8.262;20;00000000000;7;          14
     10;25;00000000003;00000000007;00000000000;00000000000;00000000001;00000000003;00000000001;00000000003;00000000002;00000000004;00000000001;00000000001;00000000000;00000000000;;   ;00000000000;00000000000;;00000000125;00000000107;00000
R;2008;010;016;;000;10860;Leonard, Justin;35 06 22;Mercedes-Benz Championship;Y ;750.000; 175,000.00;008;  T8;073;011;068;006;069;006;069;008;000;000;000;000;00000000068;00000000279;   00000000004;8;       8.262;8;00000000000;7;          18
5.083;         14;14;00000000007;00000000009;00000000001;00000000001;00000000002;00000000005;00000000002;00000000003;00000000005;00000000005;00000000002;00000000002;00000000002;00000000002;1;          25.417;00000000002;00000000001;
R;2008;010;016;;000;24781;Mahan, Hunter;25 07 20;Mercedes-Benz Championship;Y ;912.500; 218,000.00;005;  T5;073;011;072;020;069;016;064;005;000;000;000;000;00000000064;00000000278;   00000000004;5;       8.262;5;00000000001;2;          18;1
00000000002;00000000003;00000000001;00000000001;00000000002;00000000002;00000000003;00000000003;00000000000;00000000002;00000000001;00000000001;00000000000;00000000001;14;   ;00000000000;00000000000;;00000000125;00000000116;00000000023;00
R;2008;010;016;;000;24490;McNeill, George;32 03 04;Mercedes-Benz Championship;Y ;210.000;  72,000.00;025; T25;074;016;069;013;072;019;075;025;000;000;000;000;00000000069;00000000290;   00000000004;25;       8.262;25;00000000000;7;
  152.333;        12;28;00000000002;00000000005;00000000002;00000000002;00000000002;00000000007;00000000000;00000000001;00000000003;00000000006;00000000002;00000000002;00000000001;00000000003;10;          19.083;00000000002;00000000000
R;2008;010;016;;000;21949;Ogilvie, Joe;33 08 29;Mercedes-Benz Championship;Y ;162.500;  65,000.00;031;  31;075;021;073;027;074;028;074;031;000;000;000;000;00000000073;00000000296;   00000000004;31;       8.262;31;00000000000;7;          11;
      15;18;00000000003;00000000007;00000000001;00000000001;00000000002;00000000006;00000000001;00000000003;00000000003;00000000004;00000000002;00000000003;00000000001;00000000003;10;          23.667;00000000003;00000000000;;0000000013
R;2008;010;016;;000;23621;Sabbatini, Rory;31 09 04;Mercedes-Benz Championship;Y ;400.000;  90,000.00;017;  17;072;008;074;021;069;019;069;017;000;000;000;000;00000000069;00000000284;   00000000004;17;       8.262;17;00000000000;7;
03;00000000004;00000000011;29;     50.500;          8;12;00000000002;00000000008;00000000000;00000000000;00000000000;00000000004;00000000000;00000000002;00000000004;00000000005;00000000000;00000000000;00000000001;00000000001;1;
R;2008;010;016;;000;06567;Singh, Vijay;44 10 15;Mercedes-Benz Championship;Y ;525.000; 130,000.00;012; T12;074;016;070;017;067;010;070;012;000;000;000;000;00000000067;00000000281;   00000000004;12;       8.262;12;00000000000;7;          14;
250;          8;2;00000000002;00000000003;00000000002;00000000002;00000000001;00000000002;00000000000;00000000000;00000000005;00000000006;00000000001;00000000001;00000000000;00000000000;;   ;00000000000;00000000000;;00000000132;0000000012
R;2008;010;016;;000;27649;Snedeker, Brandt;27 00 29;Mercedes-Benz Championship;Y ;650.000; 155,000.00;010; T10;071;006;069;004;070;006;070;010;000;000;000;000;00000000069;00000000280;   00000000004;10;       8.262;10;00000000000;7;
00014;6;     92.833;        13;15;00000000005;00000000007;00000000003;00000000003;00000000002;00000000004;00000000000;00000000001;00000000004;00000000005;00000000004;00000000004;00000000000;00000000000;;   ;00000000000;00000000000;;00000
R;2008;010;016;;000;21528;Stenson, Henrik;31 09 01;Mercedes-Benz Championship;Y ;.000;  67,500.00;028; T28;076;026;075;028;073;031;068;028;000;000;000;000;00000000068;00000000292;   00000000004;28;       8.262;28;00000000000;7;          13;
001;00000000001;00000000001;00000000003;00000000004;00000000006;00000000003;00000000005;00000000000;00000000000;00000000001;00000000002;6;           22.417;00000000002;00000000000;;00000000128;00000000107;00000000022;00000000006;0000000
R;2008;010;016;;000;06527;Stricker, Steve;40 10 14;Mercedes-Benz Championship;Y ;2700.000; 630,000.00;002;    2;073;011;069;011;068;006;064;002;000;000;000;000;00000000064;00000000274;   00000000004;1;       8.262;1;00000000001;2;           2
00000000005;00000000009;14;     42.250;          7;7;00000000003;00000000003;00000000000;00000000000;00000000001;00000000003;00000000000;00000000001;00000000004;00000000004;00000000000;00000000000;00000000000;00000000000;;   ;00000000000;
R;2008;010;016;;000;02239;Verplank, Scott;43 05 28;Mercedes-Benz Championship;Y ;362.500;  86,000.00;018; T18;073;011;070;013;072;019;071;018;000;000;000;000;00000000070;00000000286;   00000000004;18;       8.262;18;00000000000;7;
0002;00000000002;00000000000;00000000003;00000000000;00000000001;00000000003;00000000004;00000000003;00000000003;00000000001;00000000001;1;          5.250;00000000001;00000000000;;00000000124;00000000106;00000000022;00000000002;0000000
R;2008;010;016;;000;27095;Watney, Nick;26 08 12;Mercedes-Benz Championship;Y ;912.500; 218,000.00;005;  T5;068;001;072;004;067;002;071;005;000;000;000;000;00000000067;00000000278;   00000000004;5;       8.262;5;00000000000;7;          18;10
0000000005;11;     25.083;          4;11;00000000002;00000000004;00000000000;00000000000;00000000000;00000000001;00000000001;00000000002;00000000002;00000000002;00000000000;00000000000;00000000000;00000000000;;   ;00000000000;00000000000;
R;2008;010;016;;000;26289;Weekley, Boo;34 05 14;Mercedes-Benz Championship;Y ;250.000;  76,000.00;023; T23;080;030;074;029;068;028;066;023;000;000;000;000;00000000066;00000000288;   00000000004;23;       8.262;23;00000000002;1;          14;
6;00000000017;30;     98.500;        12;17;00000000003;00000000009;00000000001;00000000001;00000000000;00000000006;00000000003;00000000005;00000000002;00000000005;00000000001;00000000001;00000000000;00000000002;14;          23.250;0000
R;2008;010;016;;000;10423;Weir, Mike;37 07 25;Mercedes-Benz Championship;Y ;1200.000; 310,000.00;004;    4;071;006;067;001;068;001;070;004;000;000;000;000;00000000067;00000000276;   00000000004;4;       8.262;4;00000000001;2;          19;7;
000002;00000000002;00000000000;00000000000;00000000001;00000000002;00000000002;00000000003;00000000003;00000000005;00000000001;00000000001;00000000002;00000000003;5;          18.583;00000000003;00000000000;;00000000121;00000000106;00000
R;2008;010;016;;000;21878;Wilson, Mark;33 02 06;Mercedes-Benz Championship;Y ;192.500;  69,000.00;027;  27;074;016;072;021;073;026;072;027;000;000;000;000;00000000072;00000000291;   00000000004;27;       8.262;27;00000000000;7;          12;
     10;29;00000000002;00000000003;00000000000;00000000001;00000000001;00000000004;00000000000;00000000003;00000000002;00000000003;00000000002;00000000002;00000000000;00000000000;;   ;00000000000;00000000001;          18;00000000133;00000
R;2008;020;006;;000;23778;Adamonis, Brad;34 11 28;Sony Open in Hawaii;Y ;132.500;  28,090.00;032; T32;066;005;068;006;070;021;074;032;000;000;000;000;00000000066;00000000278;   00000000004;32;- 1.116;32;00000000000;24;          15;24;
015;00000000025;34;    182.167;        19;51;00000000004;00000000009;00000000001;00000000001;00000000006;00000000010;00000000002;00000000004;00000000007;00000000010;00000000000;00000000001;00000000005;00000000009;32;          82.667;00
R;2008;020;006;;000;10885;Allenby, Robert;36 06 01;Sony Open in Hawaii;Y ;.000;;999; CUT;070;079;071;088;000;000;000;000;000;000;000;000;00000000070;00000000141;   00000000002;999;       .310;999;00000000000;999;          5;999;          2
0000006;00000000011;999;     37.917;          7;999;00000000001;00000000002;00000000001;00000000001;00000000001;00000000004;00000000000;00000000000;00000000004;00000000006;00000000001;00000000001;00000000003;00000000005;999;          35
R;2008;020;006;;000;23540;Allred, Jason;27 09 07;Sony Open in Hawaii;Y ;73.200;  15,518.40;045; T45;069;053;069;048;074;067;068;045;000;000;000;000;00000000068;00000000280;   00000000004;45;- 1.116;45;00000000000;24;          15;24;
000000027;37;    198.333;        25;26;00000000010;00000000017;00000000001;00000000002;00000000002;00000000005;00000000003;00000000008;00000000005;00000000008;00000000006;00000000006;00000000004;00000000006;13;          42.583;0000000
R;2008;020;006;;000;06004;Ames, Stephen;43 08 16;Sony Open in Hawaii;Y ;535.714; 113,571.43;010; T10;070;079;068;048;065;014;071;010;000;000;000;000;00000000065;00000000274;   00000000004;10;- 1.116;10;00000000000;24;          15;24;
```

Screenshot of ShotLink Data after Delimiter (.CSV File)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	Tournament ID	Player Number	Player Name	Player Age	Event Nan	FedExCup	Money	Finish Pos	Made Cut	Finish Position(text)	Round 1 S	Round 1 P	Round 2 S	Round 2 P	Round 3 S	Round 3 P	Round 4 S	Round 4 P	Lowest Rc	Round 1 +	Total Strol	Total Roui	Stroke Av	Scoring A\	Scoring A
2	30	32448	Hahn, James	33	CIMB Clas	100	252,000.00	6	Yes	6	70	33	65	12	64	4	68	6	64	135	267	4	6	9.523	6
3	40	27064	Vegas, Jhonattan	31	Sandersor	62	154,570.00	4	Yes	T4	66	6	67	2	70	3	69	4	66	133	272	4	4	5.957	4
4	100	33122	Knox, Russell	30	Hyundai T	44	66,000.00	27	Yes	27	71	19	72	21	69	20	72	27	69	143	284	4	27	12.77	27
5	100	27649	Snedeker, Brandt	35	Hyundai T	162.5	381,000.00	3	Yes	T3	67	3	72	13	65	4	67	3	65	139	271	4	3	12.77	3
6	430	23353	Henry, J.J.	41	Travelers	0		999	No	CUT	71	94	69	74	0	0	0	0	69	140	140	2	999	-0.289	999
7	20	33122	Knox, Russell	30	Shriners H	31.5	28,160.00	38	Yes	T38	69	60	70	48	72	64	66	38	66	139	277	4	38	3.683	38
8	30	32839	Matsuyama, Hideki	23	CIMB Clas	110	280,000.00	5	Yes	5	65	2	66	3	68	4	67	5	65	131	266	4	5	9.523	5
9	50	24461	Aiken, Thomas	32	World Gol	16	47,500.00	54	Yes	T54	70	42	71	37	73	54	72	54	70	141	286	4	54	5.684	54
10	100	46440	Kaufman, Smylie	24	Hyundai T	55	108,666.67	15	Yes	T15	70	12	69	13	71	15	68	15	68	139	278	4	15	12.77	15
11	100	36689	Koepka, Brooks	25	Hyundai T	162.5	381,000.00	3	Yes	T3	69	7	68	8	63	2	71	3	63	137	271	4	3	12.77	3
12	120	35451	Pinckney, Scott	26	CareerBui	51	75,897.15	17	Yes	T17	69	51	67	24	67	19	69	17	67	136	272	4	17	8.001	17
13	420	26329	Oosthuizen, Louis	33	PGA Cham	44	75,636.36	22	Yes	T22	68	9	70	22	70	30	68	22	68	138	276	4	22	-1.625	22
14	460	10809	Furyk, Jim	46	Wyndham	67.5	134,400.00	10	Yes	T10	66	8	64	2	67	3	70	10	64	130	267	4	10	4.909	10
15	460	8075	Kelly, Jerry	49	Wyndham	44	45,665.45	22	Yes	T22	67	22	68	26	65	14	70	22	65	135	270	4	22	4.909	22
16	30	29518	Steele, Brendan	32	CIMB Clas	162.5	406,000.00	3	Yes	T3	67	8	63	2	66	1	68	3	63	130	264	4	3	9.523	3
17	40	36699	Rodgers, Patrick	23	Sandersor	26.25	41,358.75	20	Yes	T20	70	63	64	4	70	8	73	20	64	134	277	4	20	5.957	20
18	40	37189	Varner III, Harold	25	Sandersor	16	14,402.18	39	Yes	T39	71	83	70	67	68	36	71	39	68	141	280	4	39	5.957	39
19	50	28237	McIlroy, Rory	26	World Gol	64.167	106,166.67	11	Yes	T11	68	16	72	27	68	21	66	11	66	140	274	4	11	5.684	11
20	100	32102	Fowler, Rickie	27	Hyundai T	110	274,000.00	5	Yes	5	69	7	67	6	69	6	67	5	67	136	272	4	5	12.77	5
21	100	28679	Gomez, Fabian	37	Hyundai T	91.667	202,333.33	6	Yes	T6	68	6	66	2	70	4	71	6	66	134	275	4	6	12.77	6
22	100	30926	Kirk, Chris	30	Hyundai T	46	70,000.00	24	Yes	T24	70	12	71	18	70	18	72	24	70	141	283	4	24	12.77	24
23	130	20572	Pampling, Rod	46	Farmers Ir	0		999	No	CUT	71	54	73	73	0	0	0	0	71	144	144	2	999	-0.577	999
24	390	29268	Burgoon, Bronson	29	Barbasol C	9.5	7,910.00	54	Yes	T54	69	40	70	56	68	36	70	54	68	139	277	4	54	6.599	54
25	420	25364	Casey, Paul	39	PGA Cham	77	233,000.00	10	Yes	T10	69	21	69	22	68	14	67	10	67	138	273	4	10	-1.625	10
26	420	28089	Day, Jason	28	PGA Cham	330	1,080,000.00	2	Yes	2	68	9	65	3	67	2	67	2	65	133	267	4	2	-1.625	2
27	460	12510	Campbell, Chad	42	Wyndham	11.5	12,376.00	57	Yes	T57	69	58	66	26	72	62	68	57	66	135	275	4	57	4.909	57
28	460	34431	Streb, Robert	29	Wyndham	34	27,751.11	33	Yes	T33	67	22	67	11	68	25	69	33	67	134	271	4	33	4.909	33
29	50	35296	Li, Hao Tong	20	World Gol	0	173,750.00	7	Yes	T7	66	8	69	4	66	3	72	7	66	135	273	4	7	5.684	7
30	270	37273	Ernst, Derek	25	Zurich Cla	29.5	28,029.16	36	Yes	T36	67	5	70	10	73	36	0	0	67	137	210	3	36	3.001	36
31	30	25364	Casey, Paul	38	CIMB Clas	46	61,600.00	24	Yes	T24	69	24	69	28	69	29	68	24	68	138	275	4	24	9.523	24
32	30	33413	Martin, Ben	28	CIMB Clas	21.5	18,223.34	47	Yes	T47	70	33	70	45	68	32	73	47	68	140	281	4	47	9.523	47
33	50	33122	Knox, Russell	30	World Gol	550	1,400,000.00	1	Yes	1	67	9	65	2	68	1	68	1	65	132	268	4	1	5.684	1
34	50	24138	Poulter, Ian	39	World Gol	39	68,000.00	30	Yes	T30	72	60	71	50	70	48	67	30	67	143	280	4	30	5.684	30
35	100	20472	Cejka, Alex	45	Hyundai T	49	78,333.33	21	Yes	T21	72	25	72	26	69	24	68	21	68	144	281	4	21	12.77	21
36	100	34360	Reed, Patrick	25	Hyundai T	300	688,000.00	2	Yes	2	65	1	69	2	67	3	69	2	65	134	270	4	2	12.77	2
37	100	25804	Watson, Bubba	37	Hyundai T	65.4	148,000.00	10	Yes	T10	69	7	68	8	73	15	67	10	67	137	277	4	10	12.77	10
38	300	25686	Dufner, Jason	39	AT&T Byrc	42.5	54,020.00	24	Yes	T24	66	18	67	16	68	13	70	24	66	133	271	4	24	3.467	24

Column	Data Type
Tournament.ID	int
Player.Number	int
Player.Name	Factor
Player.Age	int
Event.Name	Factor
FedExCup.Points	num
Money	Factor
Finish.Position.numeric.	int
Finish.Bracket	Factor
Made.Cut	Factor
Finish.Position.text.	Factor
Round.1.Score	int
Round.1.Pos	int
Round.2.Score	int
Round.2.Pos	int
Round.3.Score	int
Round.3.Pos	int
Round.4.Score	int
Round.4.Pos	int
Lowest.Round	int
Round.1...2.Strokes	int
Total.Strokes	int
Total.Rounds.	int
Stroke.Average..Rank.	int
Scoring.Avg.Total.Adjustment.	num
Scoring.Avg.Total.Adjustment...Rank.	int
Eagles	int
Eagles..Rank.	int
Birdies	int
Birdie.	num
Birdies..Rank.	int
Pars	int

Par.	num
Bogeys	int
Bogey.	num
Bogeys..Rank.	int
Doubles	int
Others	int
Total.Holes.Over.Par	int
Bogey.Avoidance.Rank	int
Birdie.or.Better.Conv.....Birdies.	int
Birdie.or.Better.Conv.....Greens.Hit.	int
Longest.Drive	int
Longest.Drive..Rank.	int
Driving.Distance.Total.Distance.	int
Driving.Distance.Total.Drives.	int
Driving.Distance..Rank.	int
Driving.Dist....All.Drives.Tot..Dist..	int
Driving.Dist....All.Drives.Rank.	int
Drives.Over.300.Yards....of.Drives.	int
Driving.Acc....Fairways.Hit.	int
Driving.Acc....Possible.Fairways.	int
Driving.Accuracy.Rank	int
Total.Driving..Rank.	int
Left.Rough.Tendency.Total.Left.Rough.	int
Right.Rough.Tendency.Total.Right.Rough.	int
App..50.75.yds.ft.	num
App..50.75.yds.attempts.	int
App..75.100.yds.ft.	num
App..75.100.yds.attempts.	int
App..100.125.yds.ft.	num
App..100.125.yds.attempts.	int
App..50.125.Yards.ft.	num
App...50.125.Yards.attempts.	int
Approaches..125.150.Yards.ft.	num
Approaches...125.150.Yards.attempts.	int

Approaches...150.175.Yards.ft.	num
Approaches...150.175.Yards.attempts.	int
Approaches...175.200.Yards.ft.	num
Approaches...175.200.Yards.attempts.	int
Approaches....200.Yards.ft.	Factor
Approaches....200.Yards.attempts.	int
App..50.75.yds.ft....Rough	num
App..50.75.yds.attempts....Rough	int
App..75.100.yds.ft....Rough	num
App..75.100.yds.attempts....Rough	int
App..100.125.yds.ft....Rough	num
App..100.125.yds.attempts....Rough	int
Approaches.50.125.Yards.ft....Rough	num
Approaches.50.125.Yards.attempts....Rough	int
Approaches.125.150.Yards.ft....Rough	num
Approaches.125.150.Yards.attempts....Rough	int
Approaches.150.175.Yards.ft....Rough	num
Approaches.150.175.Yards.attempts....Rough	int
Approaches.175.200.Yards.ft....Rough	num
Approaches.175.200.Yards.attempts....Rough	int
Approaches..200.Yards.ft....Rough	num
Approaches....200.Yards.attempts....Rough	int
GIR.	num
GIR.Missed.	num
Total.Holes.Played	int
Greens.Missed.in.Regulation	int
Total.Greens.in.Regulation	int
GIR.Rank	int
Total.Distance.ft..Prox.to.Hole	Factor
X..of.Attempts.Prox.to.Hole	int
Proximity.to.Hole..Rank.	int
Fairway.Prox.attempts.	int
Fairway.Prox.distance.in.ft.	Factor
Fairway.Prox..Rank.	int

Rough.Prox.attempts.	int
Rough.Prox.distance.in.ft.	Factor
Rough.Prox..Rank.	int
Left.Rough.Prox.attempts.	int
Left.Rough.Prox.distance.in.ft.	num
Right.Rough.Prox.attempts.	int
Right.Rough.Prox.distance.in.ft.	Factor
Going.for.Green.attempts.	int
Going.for.Green.non.attempts.	int
Going.for.the.Green.successes.	int
Scrambling.Par.or.Better	int
Scrambling.	num
Scrambling.Missed.GIR	int
Scrambling..Rank.	int
Scrambling.Proximity..Total.Distance.	num
Scrambling.Proximity....of.shots.	int
Scrambling.Proximity..rank.	int
Scrambling.from.the.Rough.successes.	int
Scrambling.from.the.Rough.attempts.	int
Scrambling.from.the.Fringe.successes.	int
Scrambling.from.the.Fringe.attempts.	int
Scrambling....30.Yards.successes.	int
Scrambling....30.Yards.attempts.	int
Scrambling.20.30.Yards.successes.	int
Scrambling...20.30.Yards.attempts.	int
Scrambling.10.20.Yards.successes.	int
Scrambling.10.20.Yards.attempts.	int
Scrambling...10.Yards.successes.	int
Scrambling....10.Yards.attempts.	int
Sand.Save.....Saves.	int
Sand.Save.....Bunkers.	int
Sand.Save..Rank.	int
Prox.to.Hole.from.Sand.Total.Distance.	num
Prox.to.Hole.from.Sand...of.Shots.	int

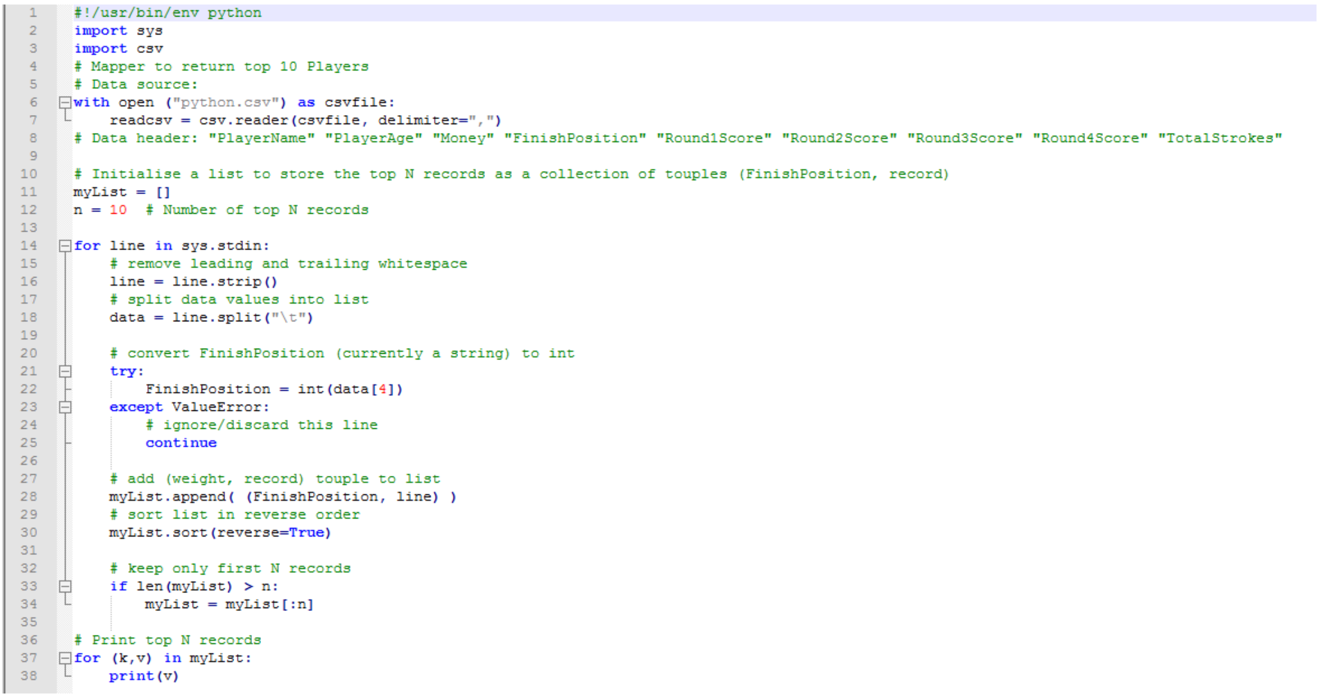
Total.Hole.Outs	int
Longest.Hole.Out.yards.	int
Overall.Putting.Avg...of.Putts.	int
Putting.Avg.GIR.Putts.	int
GIR.Putt.AVG	num
One.Putt.....of.One.Putts.	int
X3.Putt.Avoid.Total.3.Putts.	int
Approach.Putt.Performance.attempts.	int
Approach.Putt.Performance.ft..	num
Avg.Distance.of.Putts.Made.Total.Distance.of.Putts.:	3221
Total.Rounds.Played	int
Putting.3..attempts.	int
Putting.3..putts.made.	int
Putting...4..attempts.	int
Putting.4..putts.made.	int
Putting.5..attempts.	int
Putting.5..putts.made.	int
Putting.6..attempts.	int
Putting.6..putts.made.	int
Putting.7..attempts.	int
Putting.7..putts.made.	int
Putting.8..attempts.	int
Putting.8..putts.made.	int
Putting.9..attempts.	int
Putting.9..putts.made.	int
Putting.10..attempts.	int
Putting.10..putts.made.	int
Putting.Inside.5...putts.made.	int
Putting.Inside.5...attempts.	int
Putting.Inside.5.Feet..Rank.	int
Putting.5....10...putts.made.	int
Putting.5....10...attempts.	int
Putting.5....10...Rank.	int
Putting.4..8..attempts.	int

Putting.4..8..putts.made.	int
Putting.4....8...Rank.	int
Putting.Inside.10..attempts.	int
Putting.Inside.10..putts.made.	int
Putting.Inside.10..rank.	int
Putting.10..15..attempts.	int
Putting...10..15..putts.made.	int
Putting.10..15..rank.	int
Putting.15.20..attempts.	int
Putting.15..20..putts.made.	int
Putting.15..20..rank.	int
Putting.20..25..attempts.	int
Putting.20..25..putts.made.	int
Putting.20..25..rank.	int
Putting..25..attempts.	int
Putting..25..putts.made.	int
Putting..25..rank.	int
Putting...10...putts.made.	int
Putting...10...attempts.	int
Putting...10...rank.	int
Total.Putts.Gained	num
Total.Rounds.Played.Putts.Gained.	int
Putts.Gained.Rank	int
TTL.SG.T2G	num
SG..T2G.Rank	int
TTL.SG.Total	num
SG..Total.Rank	int
OTT.SG.Avg.	num
OTT.SG.Rank	int
APP.SG.Avg.	num
APP.SG.Rank	int
ARG.SG.Avg.	num
ARG.SG.Rank	int

Appendix D – Sample Code

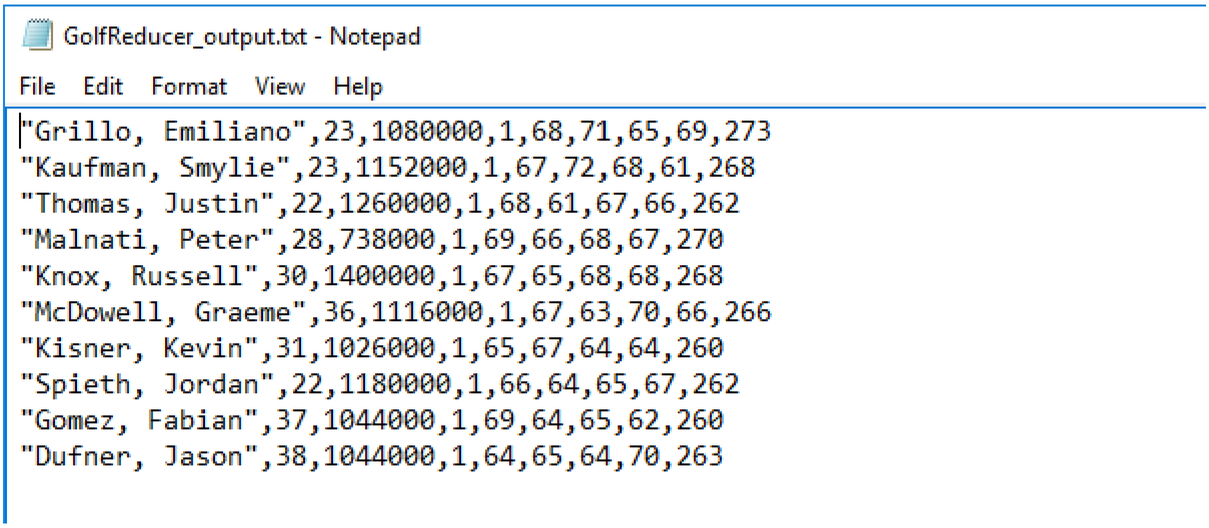
Python Script

Mapper and Reducer to output Top 10 Players for Finish Position



```python
#!/usr/bin/env python
import sys
import csv
# Mapper to return top 10 Players
# Data source:
with open ("python.csv") as csvfile:
    readcsv = csv.reader(csvfile, delimiter=",")
# Data header: "PlayerName" "PlayerAge" "Money" "FinishPosition" "Round1Score" "Round2Score" "Round3Score" "Round4Score" "TotalStrokes"

# Initialise a list to store the top N records as a collection of touples (FinishPosition, record)
myList = []
n = 10  # Number of top N records

for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # split data values into list
    data = line.split("\t")

    # convert FinishPosition (currently a string) to int
    try:
        FinishPosition = int(data[4])
    except ValueError:
        # ignore/discard this line
        continue

    # add (weight, record) touple to list
    myList.append( (FinishPosition, line) )
    # sort list in reverse order
    myList.sort(reverse=True)

    # keep only first N records
    if len(myList) > n:
        myList = myList[:n]

# Print top N records
for (k,v) in myList:
    print(v)
```

Output of Mapper and Reducer (Python)



```
"Grillo, Emiliano",23,1080000,1,68,71,65,69,273
"Kaufman, Smylie",23,1152000,1,67,72,68,61,268
"Thomas, Justin",22,1260000,1,68,61,67,66,262
"Malnati, Peter",28,738000,1,69,66,68,67,270
"Knox, Russell",30,1400000,1,67,65,68,68,268
"McDowell, Graeme",36,1116000,1,67,63,70,66,266
"Kisner, Kevin",31,1026000,1,65,67,64,64,260
"Spieth, Jordan",22,1180000,1,66,64,65,67,262
"Gomez, Fabian",37,1044000,1,69,64,65,62,260
"Dufner, Jason",38,1044000,1,64,65,64,70,263
```

Sample R Code

```
## Luke Chamberlain
## x13521763 National College of Ireland
setwd("H:/Project Data")
revent <- read.csv(file="revent(New).csv",head=TRUE,sep=",")
Driving.Distance <- read.csv(file="Top 10 driving distance.csv",head=TRUE,sep=",")
Driving.Accuracy <- read.csv(file="Top 10 accuracy.csv",head=TRUE,sep=",")

summary(revent)
str(revent, list.len=ncol(df))

library(ggplot2)

## Data Prep
revent$Money = as.integer(gsub(",", "",(revent$Money)))
MissedCut <- revent[which(revent$Finish.Position.numeric. == 999 ), ]
MadeCut <- revent[which(revent$Finish.Position.numeric. != 999), ]
No3Rounds <- revent[which(revent$Total.Holes.Played != 54), ]
No3Rounds.Cut <- No3Rounds[which(No3Rounds$Finish.Position.numeric. != 999), ]

### Par, Birdies, Bogeys
summary(revent$Birdie.)
summary(revent$Birdies)
summary(revent$Bogey.)
summary(revent$Bogeys)
summary(revent$Par.)
summary(revent$Pars)

sum(revent$Pars)

sum(revent$Birdies, na.rm = TRUE)

sum(revent$Eagles, na.rm = TRUE)

sum(revent$Bogeys, na.rm = TRUE)

sum(revent$Doubles, na.rm = TRUE)

sum(revent$Others, na.rm = TRUE)

plot(revent$Birdie., revent$Total.Strokes)


slices <- c(199288, 60482,1434, 48055, 5673, 842)

lbls <- c("Pars", "Birdies","Eagles", "Bogeys", "Doubles", "Others")

pct <- round(slices/sum(slices)*100)

lbls <- paste(lbls, pct) # add percents to labels

lbls <- paste(lbls,"%",sep="") # ad % to labels

pie(slices,labels = lbls, col=rainbow(length(lbls)),

    main="Pie Chart of Countries")

plot(slices)


###Alex Cejka vs Json Day, Birdie / Bogey Ratio

JasonDay <- (revent[which(revent$Player.Name == "Day, Jason"),])

AlexCejka <- (revent[which(revent$Player.Name == "Cejka, Alex"),])


sum(JasonDay$Money, na.rm = TRUE)

sum(AlexCejka$Money, na.rm = TRUE)


boxplot(JasonDay$GIR., main = "Jason Day & Alex Cejka", xlab = "% of Greens
in Regulation",

        AlexCejka$GIR.)
```

```
###Birdies vs.Player Age
mean(revent$Player.Age, na.rm = TRUE)
mean(revent$Birdies, na.rm = TRUE)
ggplot(revent, aes(x = Birdies, fill = factor(Player.Age))) +
  geom_bar() +
  xlab("Birdies") +
  ylab("Total Count") +
  labs(fill = "Age")

###Round Score
summary(revent$Round.1.Score)
summary(revent$Round.2.Score)
summary(revent$Round.3.Score[revent$Round.3.Score!=0])
summary(revent$Round.4.Score[revent$Round.4.Score!=0])

boxplot(revent$Round.1.Score, revent$Round.2.Score,
        revent$Round.3.Score[revent$Round.3.Score!=0],
        revent$Round.4.Score[revent$Round.4.Score!=0])


(revent[which(revent$Round.4.Score == 58),]$Player.Name)

summary(MadeCut$Round.1...2.Strokes)
summary(MissedCut$Round.1...2.Strokes)

Round.1.Importance68.WithoutCUT<-   (revent[which(revent$Round.1.Score   <=
68),])
```

```
summary(Round.1.Importance.WithoutCUT$Finish.Position.numeric. !=999)


Round.1.Importance74.WithoutCUT<-   (revent[which(revent$Round.1.Score   >=
74),])

summary(Round.1.Importance74.WithoutCUT$Finish.Position.numeric. !=999)


Round.2.Importance74.WithoutCUT<-   (revent[which(revent$Round.2.Score   >=
74),])

summary(Round.2.Importance74.WithoutCUT$Finish.Position.numeric. !=999)


Round.2.Importance.WithoutCUT<-    (revent[which(revent$Round.2.Score    <=
68),])

summary(Round.2.Importance.WithoutCUT$Finish.Position.numeric. !=999)


Round.1.Importance<- (MadeCut[which(MadeCut$Round.1.Score <= 68),])

summary(Round.1.Importance$Finish.Position.numeric.)


Round.2.Importance<- (MadeCut[which(MadeCut$Round.2.Score <= 68),])

summary(Round.2.Importance$Finish.Position.numeric.)


Round.3.Importance<- (MadeCut[which(MadeCut$Round.3.Score <= 68),])

summary(Round.3.Importance$Finish.Position.numeric.)


Round.4.Importance<- (MadeCut[which(MadeCut$Round.4.Score <= 68),])

summary(Round.4.Importance$Finish.Position.numeric.)


Round.1.Importance.74<- (MadeCut[which(MadeCut$Round.1.Score >= 74),])

summary(Round.1.Importance.74$Finish.Position.numeric.)
```

```
Round.2.Importance.74<- (MadeCut[which(MadeCut$Round.2.Score >= 74),])

summary(Round.2.Importance.74$Finish.Position.numeric.)


Round.3.Importance.74<- (MadeCut[which(MadeCut$Round.3.Score >= 74),])

summary(Round.3.Importance.74$Finish.Position.numeric.)


Round.4.Importance.74<- (MadeCut[which(MadeCut$Round.4.Score >= 74),])

summary(Round.4.Importance.74$Finish.Position.numeric.)
```

Sample Model R Code

```
setwd("H:/Project Data")

## Read Data

Ten17 <- read.csv(file="2010-2017.csv",head=TRUE,sep=",")

Masters2017 <- read.csv(file="Masters 2017.csv",head=TRUE,sep=",")

Masters <- Ten17[which(Ten17$Event.Name == "Masters Tournament"), ]

summary(Ten17$Event.Name)

summary(Ten17$Cut.Value)

Ten17$Made.Cut = as.numeric(Ten17$Made.Cut)


fit1 <- lm(Finish.Position.numeric. ~ Birdies + Pars + Bogeys + Doubles + 
Total.Greens.in.Regulation + Driving.Acc....Fairways.Hit., data = Masters)

summary(fit1)

predict(fit1)

predict1 <- predict(fit1)

merge <- data.frame(predict1, Masters [1:486,14])

MastersPredict <- predict(fit1, Masters2017)

merge1 <- data.frame(MastersPredict, Masters2017)
```

```
##Phil Mickelson

table(Masters$Player.Name == "Woosnam, Ian")

PhilMasters <- Masters[which(Masters$Player.Name == "Mickelson, Phil"), ]


mean(PhilMasters$Total.Strokes)

mean(PhilMasters$Birdies)

mean(PhilMasters$Pars)

mean(PhilMasters$Bogeys)

mean(PhilMasters$Doubles, na.rm = TRUE)

mean(PhilMasters$Total.Greens.in.Regulation)

mean(PhilMasters$Driving.Acc....Fairways.Hit.)


PhilMasters2017 <- data.frame(Birdies = 13.42, Pars = 37.285, Bogeys =
8.285, Doubles = 2.25, Total.Greens.in.Regulation = 41.85,
Driving.Acc....Fairways.Hit. = 30.285)

predict(fit1, PhilMasters2017)


#GarciaMasters2017

GarciaMasters <- Masters[which(Masters$Player.Name == "Garcia, Sergio"), ]


mean(GarciaMasters$Cut.Value)

mean(GarciaMasters$Birdies)

mean(GarciaMasters$Pars)

mean(GarciaMasters$Bogeys)

mean(GarciaMasters$Doubles, na.rm = TRUE)

mean(GarciaMasters$Total.Greens.in.Regulation)

mean(GarciaMasters$Driving.Acc....Fairways.Hit.)
```

```
GarciaMasters2017 <- data.frame(Cut.Value = 1.857, Birdies = 14.285, Pars =
37.714, Bogeys = 13.142, Doubles = 2, Total.Greens.in.Regulation = 42.428,
Driving.Acc....Fairways.Hit. = 35.142)


predict(fit1, GarciaMasters2017)
```

Appendix E – Reflective Journals

Reflective Journal

Student name: Luke Chamberlain x13521763

Programme (e.g., BSc in Computing): BSc in Technology Management

Month: September

My Achievements

This month, I came up with my idea for my final year project in Data Analytics. The idea is to predict trends and patterns in the success of performance of a professional golfer and make a comparison to betting odds/results. The idea came about as a golf enthusiast to try and see if I could analyse a large dataset that interest's me. The idea was a result of a lot of brainstorming and research. In my research I found that there weren't many other projects/regressions done on this subject. I feel this is an opportunity to highlight my research and results once complete as there has not been much research done in this area.

This Month, I also pitched my idea to 3 supervisors. This is a "Vetting" Process used to eliminate the bad project ideas. The students whose ideas are rejected are given a mandatory project idea by the project supervisors. My pitch was successful and got approval from each of the three supervisors. This meant I could continue on and pursue my idea.

My Reflection

Upon reflection of the first month of this project, I learned that there is a lot of planning and organisation needed to be done in order to complete a successful project.

Intended Changes

Next month, I plan to add more originality to my current idea to make it completely individual to any idea that has been done before. I also plan to gain access to the database I will be using "Shotlink." This database does not freely give out information but does provide the data for those who decide to use for research methods.

Supervisor Meetings

A supervisor has not been assigned yet.

Reflective Journal

Student name: Luke Chamberlain x13521763

Programme (e.g., BSc in Computing): BSc in Technology Management

Month: October

My Achievements

During October, there were several tasks which I achieved. I completed my project proposal which outlined all my plans which I hope to implement into my project over the coming months. It featured the main objectives of the project and a background into the project and why I chose it. I also discussed some of the technical details and special resources I will require to complete a successful project. Another achievement was gaining access to the database I required to access the relevant data I needed to complete my project. "ShotLink" is the database I will be using, and to gain access I had to contact the people who had authority to grant me access to the database as it is not a public database and is used for academic research purposes. I also had to complete a questionnaire which highlighted what were my purposes and intent to use the data for.

My Reflection

Upon reflection, I learned that there is a lot more research than first expected, some of the research involved was analysing other dissertations and projects like the one I hope to complete in order to gain an understanding of what may need to be partaken.

Intended Changes

Some changes I plan to make in the coming month is to stay true to my project proposal and its plan guideline. I also hope to increase some of my technical skills required for the project such as R and Python. I also hope to add more originality to my project idea as I believe this can be ever improved as I gradually complete my project.

Supervisor Meetings

Eugene O'Loughlin has been assigned as my supervisor during the month of October. We met once during October to discuss general progress of this project.

Reflective Journal

Student name: Luke Chamberlain x13521763

Programme: BSc in Technology Management

Month: December

My Achievements

During December, I prepared for the mid-point presentation. This was a presentation based on all the work completed to date thus far. It featured discussing the requirements of the project, future plans, struggles to date, key aspects of the project etc. The presentation took place in mid-December. The presentation was made in front of two college lecturers who had relevant experience with the chosen subject, one been my supervisor. The presentation did not go to plan due to lack of preparation, clarity and shortage of supervisor meetings.

Exam prep also started in this month.

My Reflection

Upon Reflection of the mid-point presentation I realised that there is a lot more work to be expected with regards to the project. More clarity is needed to progress forward with a successful project. One of my main downfalls in a poor mid-point presentation was failing to meet with my supervisor. This put me at a disadvantage as I could have easily had more clarification of what was to be expected from the presentation if I had met with my supervisor more frequently.

Intended Changes

To Organise a lot more meeting with my supervisor.

To provide a clearer understanding of the overall goal of the project.

To set a new plan in place.

To begin project coding.

Supervisor Meetings

Mid-point presentation.

Reflective Journal

Student name: Luke Chamberlain x13521763

Programme: BSc in Technology Management

Month: January

My Achievements

The majority of January was taken up with exams. Once the exams were completed I set out a new timeframe plan for the project. I also began to start working with the dataset at hand. This involved downloading all the relevant software.

My Reflection

Upon Reflection of this Month, I feel that my new project timeline will allow me to successfully complete my project at a high standard within the required timeline. I will need to enhance some of my analytical skills to further progress.

Intended Changes

I intend to make several small changes with regards to some of the questions to be answered with data in the final dissertation.

Supervisor Meetings

I met with my supervisor once in January to discuss the mid-point presentation and where I went wrong. We also discussed plans moving forward.

Reflective Journal

Student name: Luke Chamberlain x13521763

Programme: BSc in Technology Management

Month: February

My Achievements

During February the data was cleansed and processed. This was quite a time consuming process. There was also some data exploration done.

My Reflection

Upon Reflection of this Month, I feel there is still plenty of work to do in order to complete the project in the required time constriction.

Intended Changes

Several variables are to be added to the data to justify some results.

Supervisor Meetings

Several meetings were held in February to discuss progress and plans moving forward.

Reflective Journal

Student name: Luke Chamberlain x13521763

Programme: BSc in Technology Management

Month: March

My Achievements

March was a busy month with preparation for exams disturbing the progress of the project. Further analysis was complete after encountering some technical problems and issues regarding R.

My Reflection

Upon Reflection of this Month, I feel there is still plenty of work to do in order to complete the project in the required time constriction.

Intended Changes

Changes are to be made to the current documentation including format changes. Some criteria changes are to be made for analysis purposes.

Supervisor Meetings

Several meetings were held in March. All work to date was displayed to Supervisor and was informed about plans moving forward.