Professional Golf Database Documentation

Brad Klassen

bradklassen@outlook.com

Contents

1 Introduction

In the age of big data, the field of sports analytics is becoming increasingly transdisciplinary, combining domain specific knowledge from sports management with the statistical and computational tools of data science. Golf is a sport that generates massive amounts of data with no shortage of opportunity for analysis. As statistics becomes more integrated into professional sports, having an analytical edge gives both athletes and teams a competitive advantage.

2 Acquired Data

In order to receive the most accurate tournament predictions, it is important to create data sets with as much information as possible. The more variables to test with the model, the greater the likelihood of an accurate prediction. Six web scraping programs have been created in order to acquire professional golf data. The six programs acquire the following data, PGA Tour Statistics, PGA Tour Scorecard data, PGA Tour Course History, PGA Tour Tournament History, Official World Golf Ranking (OWGR) data & LPGA Tour Statistics. The programming language Python was used to create the web-scraping programs. A few of the main libraries used to acquire, manipulate and analyze the data include, BeautifulSoup, Selenium, Pandas, NumPy, and many others.

2.1 PGA Tour Statistics

There are several hundreds of statistics recorded on the PGA Tour that are used to analyze the performance of athletes at each tournament. The statistics can be divided into the following sub-categories, Off the Tee, Approach the Green, Around the Green, Putting, Scoring, Streaks, Money/Finishes and Points/Rankings. An example of the statistics include, Driving Distance, Driving Accuracy Percentage, Club Head Speed, Ball Speed, Greens in Regulation Percentage, One-Putt Percentage, and many more. Every observation from each of the sub-categories has been scraped from the PGA Tour Statistics website. Within each statistic there are variables that help to indicate the players performance, including the rank of the player in the given statistic for the current week, and the value of the statistic of interest. The PGA Tour offers the data in two formats, either the performance year-to-date or the performance at each tournament. The PGA Tour Statistics data set contains data for each tournament dating back to 1980. The tournament data makes it easy to analyze week-to-week differences in statistics which is optimal for making predictions.

2.1.1 Example

The PGA Tour Statistics data set contains tournament data from 1980 to the most recent week. The data set is updated at the conclusion of each tournament. The data typically consists of the average or sum of the players statistics at the given tournament. Below is an example of the format for the PGA Tour Statistics data set. The example consists of four players, from four years, and four different statistics from four tournaments.

Player Name	Date	Tournament	Statistic	Variable	Value
Jack Nicklaus	1980-06-15	U.S. Open Championship	Driving Distance	AVERAGE	283.0
Jack Nicklaus	1980-06-15	U.S. Open Championship	Driving Distance	RANK THIS WEEK	3
Tom Watson	1982-08-08	PGA Championship	Putts Per Round	AVERAGE	29.00
Tom Watson	1982-08-08	PGA Championship	Putts Per Round	RANK THIS WEEK	25
Tiger Woods	2000-04-09	Masters Tournament	Par 5 Performance	STATUS	-12
Tiger Woods	2000-04-09	Masters Tournament	Par 5 Performance	RANK THIS WEEK	1
Rory McIlroy	2019-03-17	THE PLAYERS Championship	SG: Off-the-Tee	AVERAGE	1.327
Rory McIlroy	2019-03-17	THE PLAYERS Championship	SG: Off-the-Tee	RANK THIS WEEK	2

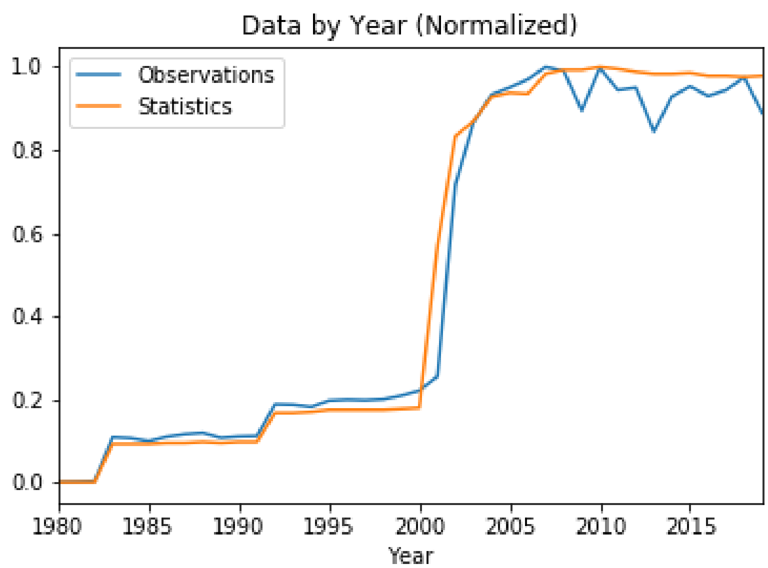
2.1.2 Yearly Data Summary

The PGA Tour has recently realized the importance of data and has increased it's data collection processes. The following table shows by year, the number of unique observations, players, dates, tournaments and statistics.

Year	Observations	Player Name	Date	Tournament	Statistic
1980	141,646	321	40	43	25
1981	143,312	334	41	43	25
1982	146,180	312	42	44	25
1983	363,968	286	42	42	63
1984	359,264	301	41	41	63
1985	344,280	293	42	42	63
1986	366,414	297	43	45	64
1987	379,270	311	44	48	64
1988	385,731	329	46	49	65
1989	361,156	304	43	45	64
1990	367,914	314	42	45	65
1991	370,094	325	43	44	65
1992	526,886	325	43	44	94
1993	524,570	337	42	43	94
1994	513,636	330	43	43	95
1995	545,174	362	43	44	97
1996	549,494	378	42	45	97
1997	547,092	350	43	45	97
1998	551,860	361	43	45	97
1999	569,912	345	44	46	98
2000	594,822	353	45	48	99
2001	664,880	355	42	46	260
2002	1,613,044	353	43	47	370
2003	1,923,360	349	44	48	385
2004	2,059,820	369	43	47	409
2005	2,095,408	382	43	47	413
2006	2,134,962	364	44	48	412
2007	2,197,374	342	44	47	432
2008	2,177,168	374	44	48	436
2009	1,980,992	357	40	44	436
2010	2,192,226	341	42	46	439
2011	2,082,540	316	40	44	437
2012	2,094,894	342	40	44	434
2013	1,877,490	329	37	40	432
2014	2,049,746	364	43	45	432

Year	Observations	Player Name	Date	Tournament	Statistic
2015	2,101,136	361	43	47	433
2016	2,052,520	351	42	46	430
2017	2,082,712	361	43	47	430
2018	2,145,686	384	44	48	429
2019	1,969,266	379	41	46	430

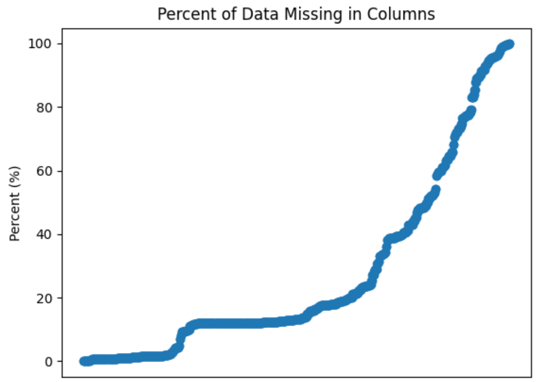
As shown by the above table and below plots, in 2001 and 2002 there appears to be a significant increase in the amount of data collected by the PGA Tour. The significant increase in data is likely due to the PGA Tour implementing Shotlink. The Shotlink system offers information on every shot taken by every player on the PGA Tour. In 2001 to 2002 the PGA Tour began to record shot-level data and from 2003 onward they have recorded quality shot-level data.

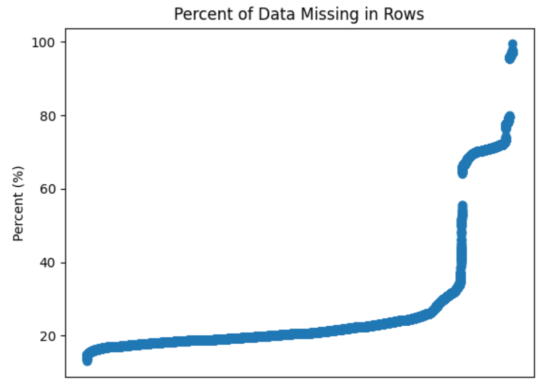


2.1.3 Missing Value Analysis

In order to get a better understanding of the data one must take a further look into the missing values that exist in the data set. Getting a better understanding of the missing values is crucial in order to determine how to deal with them.

After analyzing the distribution of the data by year, we have decided to remove any data prior to 2003 from the analysis. The initial wide formatted data set for the 2003 to 2019 seasons, consists of 54,804 rows and 887 columns. In the wide format, the rows consist of players and the tournaments they played on tour. Whereas, the columns are the statistics from the PGA Tour website. Below are graphs showing the percentage of missing data in columns (left) and percent of missing data in rows (right).





Statistic	Value
Count	887
Mean	27.2164
Standard Deviation	28.6273
Minimum	0
25%	10.9773
50%	13.1432
75%	40.1449
Maximum	99.8522

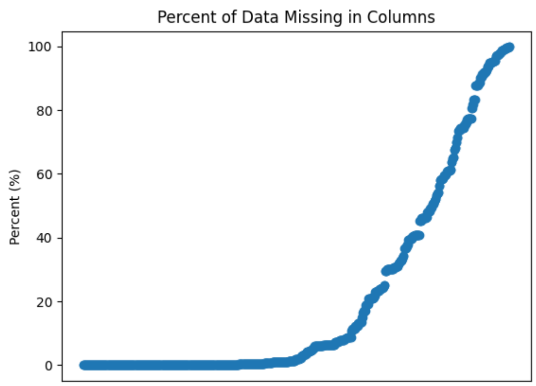
Statistic	Value
Count	54804
Mean	27.2164
Standard Deviation	17.2882
Minimum	13.0778
25%	18.7148
50%	20.5186
75%	24.3517
Maximum	99.4363

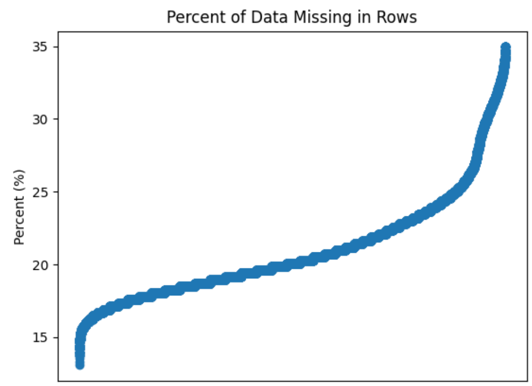
As shown above, there appears to be a large amount of missing data. In the statistic "Official World Golf Ranking", players may be assigned values for the statistic nearly every year regardless of whether they compete or not. A players Official World Golf Ranking (OWGR) points stay on the players tally for two years.[1] Therefore, it is possible that a very dominate player may still be on the OWGR leader board even if they do not play on the PGA Tour for over a year. Since the data is in a wide format, if there is only one variable with a value, then the values of every other variable in the row will be null. The "Official World Golf Ranking" statistic taking into consideration 34 professional golf tours. Many of these tours are not as prominent as the PGA Tour and often lack data. As a result, nearly all of the statistics for an athlete for the given year will be null.

Observation Threshold

As shown by the above graph for missing data in rows, there appears to be an elbow at approximately 35%. In order to reduce the number of missing values, a threshold of 35% percent missing data has been set. After the threshold has been set, there remains 48,060 rows and 887 columns. Therefore, 6,744 rows have been excluded from the data. After setting the missing value threshold on rows, the data is as follows.

[1] **official˙world˙golf˙ranking**.





Statistic	Value
Count	887
Mean	20.9737
Standard Deviation	30.8252
Minimum	0
25%	0.0042
50%	2.0204
75%	33.8618
Maximum	99.8315

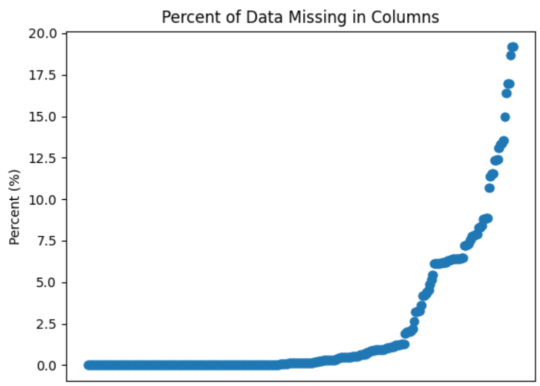
Statistic	Value
Count	48060
Mean	20.9737
Standard Deviation	3.6583
Minimum	13.0778
25%	18.4893
50%	20.0676
75%	22.5479
Maximum	34.9493

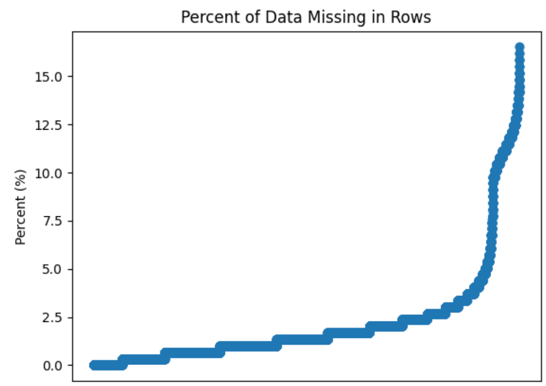
Statistic Threshold

Setting the threshold on the rows resulted in a slight change in the missing data in the columns. After analyzing the graph for missing values in columns there appears to be an elbow at approximately 20%. In addition to analyzing the graph, it is important to look at the statistics that will be excluded as a result of a column threshold. After analyzing both the graph and columns, a column threshold has been set on 20% missing data. After setting both a row threshold and column threshold, the data is as follows.

Statistic	Value
Count	593
Mean	2.1615
Standard Deviation	3.9744
Minimum	0
25%	0
50%	0.1498
75%	2.0204
Maximum	19.1760

Statistic	Value
Count	48060
Mean	2.1615
Standard Deviation	2.7089
Minimum	0
25%	0.6745
50%	1.3491
75%	2.3609
Maximum	16.5261





The final data set now consists of 48,060 rows and 593 columns. Where rows are the player in a given tournament, and columns are the variables. There are now 1,154 players in the data set.

2.1.4 Missing Value Imputation

Many supervised machine learning models do not allow missing data when training the model. Therefore, the missing data was imputed using two of the following methods.

K-Nearest Neighbour (KNN)

The KNN algorithm finds the 5 nearest neighbours (grouped by player) of the missing value and fills in the value with the mean of its neighbours. If there are less than 5 neighbours the algorithm will impute the missing value as the the mean of its neighbours. In the occasion that a player does not have a value for the given statistic for any tournaments, the missing value will remain blank, as it does not have a non-null neighbour.

Average of Statistic for Given Week

To remove missing values entirely, a second imputation method has been introduced. The missing value will be imputed using the mean of the statistic for the field for the given week. After completing this second imputation method there should no longer be missing values in the data set.

2.1.5 Lagging Features

Due to the time series structure of the data set, the variables will be lagged in order to make predictions for upcoming events. The features have been lagged for the previous 3 weeks. Each statistic has been repeated four times to take into account the current week and the three previous weeks. For example, a statistic in the original data set is “Official Money - (RANK THIS WEEK)” after lagging the statistic there are now four examples of this statistic to take into account. The original statistic now consists of the following four features, “Official Money - (RANK THIS WEEK) (Week t-0)’”, “Official Money - (RANK THIS WEEK) (Week t-1)”, “Official Money - (RANK THIS WEEK) (Week t-2)” and “Official Money - (RANK THIS WEEK) (Week t-3)”. When making predictions all features containing “(Week t-0)” have been removed from the training data. Therefore, the Week t-1, t-2 and t-3 features are used to predict the output variable for Week t-0. Lagging the variables reintroduced missing values into the data set for the first couple weeks of each player competed on tour. For example, when predicting performance of Tiger Woods at his second tournament, his week t-1 statistics will be populated, but week t-2 and t-3 will not be populated as he has only played one event on tour. Any observations with missing values at this stage are removed. Therefore, the first two records for each athlete on tour have been removed

2.1.6 Target Variable

The target variable used to predict performance is “Official Money - (RANK THIS WEEK)”. This statistic offers the ranking of players at a given tournament. A unique benefit of this statistic is we are able to use the “Official Money - (MONEY)” which accounts for the money earned by the player and not the ranking of the player. This feature is interesting because the more important events typically offer the opportunity to win more money. Therefore, if a player wins a larger amount of money the model can realize that the tournament was important. In a way this accounts for the strength of the field also, as events with larger purses typically have stronger fields. One slight downfall in using this feature as the target variable, is that amateur golfers can not win money in the events they compete it. Many tournaments allow a couple of amateurs to compete in the event, with the exception that they can not win any money. Therefore, if an amateur places well at a tournament we will be unaware because it will not count to their money count.

2.1.7 Survivorship Bias

A known problem with the tournament only format is that the data recorded is only for the athletes that made the cut. This introduces potential survivorship bias to the data. Survivorship bias is a sample of selection bias where the data only consists of records that ”survived” and not records that failed to meet the requirement. Since the data does not include records of athletes that missed the cut, the bias is introduced.

2.2 PGA Tour Scorecard Data

As mentioned in the previous section, the PGA Tour Statistics data suffers from survivorship bias, which can be a problem for machine learning models. Using this data the models would expect the players to perform better than they likely would. It would not penalize players with large variance as their poor tournaments would not be included in the data. To combat this bias, the PGA Tour offers scorecard data recording scoring by hole and round. Each round has a summary of a few of the most important statistics for evaluating performance. This data offers statistics for players even if they miss the cut at an event. An example of the scorecard data for a player that made the cut is as follows, Jon Rahm at the 2020 Memorial Tournament. An example for a player that missed the cut is as follows, Rickie Fowler at the 2020 PGA Championship. As shown in Rickie's statistics, player's that missed the cut do not receive a ranking for each statistic. However, the ranking can be calculated using the values from the total column of all athletes that competed in the event. The tournament statistics section of the scorecard data has been acquired. The hole-by-hole data will be acquired once there is an appropriate use case.

This data is extremely valuable for fitting the machine learning models as it does not suffer from survivorship bias. The data includes nineteen statistics which is significantly less than the approximately four-hundred statistics included in the PGA Tour Statistics data. The decrease in the number of statistics reduces the chances of overfitting the model. Many of the statistics in the PGA Tour Statisitcs data set are not a key indicator of success at upcoming tournaments and as a result may allow the model to fit on error. This will be the primary data set used to fit the machine learning models.

2.2.1 Example

The PGA Tour scorecard data set contains tournament data from 1980 to the most recent week. The data set is updated at the conclusion of each tournament. The data set consists of the player name, date and tournament as well as, many statistics rating the players performance at the given tournament. Below is an example of the format for the PGA Tour scorecard data set. The example consists of three players from three years with a sample of a few of the many important statistics. The statistics in the data set are, Player Name, Year, Date, Tournament, Course, 3+ Bogeys, Birdies, Bogeys, Double Bogeys, Driving Accuracy, Driving Distance, Eagles, Greens in Regulation, Longest Drive, Pars, Putts Per GIR, Sand Saves, Scrambling, SG: Approach to the Green, SG: Around the Green, SG: Off the Tee, SG: Putting, SG: Tee to Green and SG: Total. These statistics are repeated for each round, as well as, their total value throughout the week and their rank in the given statistic. The format of the column names is 'statistic' - ('variable'). When the variable is a number of 1 through 4, this indicates the round from which the statistic is from.

Player Name	Year	Tournament	Course	Driving Accuracy - (1)	Birdies - (Rank)	SG: Off the Tee - (Total)
Tiger Woods	2019	U.S. Open	Pebble Beach Golf Links	71.43	22	0.655
Tiger Woods	2019	Farmers Insurance Open	Torrey Pines GC (South)	50	17	0.632
Paul Casey	2018	Travelers Championship	TPC River Highlands	71.43	10	-1.496
Paul Casey	2018	Genesis Open	Riviera Country Club	64.29	23	3.857
Cameron Champ	2020	PGA Championship	TPC Harding Park	50	5	6.299
Cameron Champ	2020	Charles Schwab Challenge	Colonial Country Club	50	19	2.703

2.2.2 Automated Collection

Python scripts have been created to automatically retrieve the scorecard data at the conclusion of each tournament. The historical data 1980 - 2019 data will be retrieved once and will not change week to week. A function has been created to identify players that have competed in the 2020 season. Every Monday the script will run to acquire data for all athletes that have competed in the current year. This data will then replace the 2020 data from the previous week. A second function has been created to retrieve the name of athletes competing in this weeks upcoming tournament. Predictions will only be made for those athletes.

2.3 PGA Tour Course History

Unlike many other professional sports, the location of events played has an integral role on how certain athletes will perform. Something that seems to be as insignificant as the type of grass used at the golf course can significantly affect how players perform, especially on the greens. The altitude of a course is very important due to the air density. As the altitude increases, the air density decreases which leads to further ball flight. Humidity and wind also play a very important part in driving distance and accuracy. Having data specific to the course and location of the tournament helps to discover the athletes that perform best in certain conditions. A players previous performance at a certain course may offer insights on how the given player may perform at the course in the future.

2.3.1 Example

The PGA Tour course history data set contains tournament data from 1980 to the most recent week. The data set is updated at the conclusion of each tournament. The data set consists of the course name, designer and location as well as, many statistics rating the players performance at the given course. Below is an example of the format for the PGA Tour course history data set. The example consists of three players from two courses each with a sample of a few of the many important statistics. The statistics in the data set are, Player ID, Player Name, Course ID, Course Number, Course Name, Course Location, Course Designer, Events Played, Total Rounds, Finished First, Finished Second, Finished Third, Finished Top Ten, Finished Top Twenty Five, Number of Made Cuts, Number of Missed Cuts, Number of Disqualifications, Number of Withdraws and Total Money. Using feature engineering, new columns have been added that analyze the players performance while taking into consideration the number of events the player has competed in at the given course.

Player Name	Course Name	Course Location	Course Designer	Events Played	Top 10 Finishes
Tiger Woods	Augusta National GC	Augusta, GA	Mackenzie & Jones Jr.	22	14
Tiger Woods	Trump National Doral	Miami, FL	Dick Wilson & Robert von Hagge	11	9
Rory McIlroy	TPC Sawgrass	Ponte Vedra Beach, FL	Pete Dye	10	4
Rory McIlroy	PGA National (Champion)	Palm Beach Gardens, FL	Tom Fazio	9	2
Justin Thomas	Plantation Course at Kapalua	Kapalua, Maui, HI	Bill Coore & Ben Crenshaw	5	3
Justin Thomas	TPC River Highlands	Cromwell, CT	Robert J. Ross & Maurice Kearney	6	1

2.4 PGA Tour Tournament History

Using data provided on the PGA Tour website, it is possible to analyze performance at each tournament. This data set provides the opportunity to combine the PGA Tour Statistics data set with the PGA Tour Course History data set. This data set is crucial for implementing the course history data into a model because it has both tournament name and course name.

2.4.1 Example

The PGA Tour tournament history data set contains tournament data from 1980 to the most recent week. The data set is updated at the conclusion of each tournament. The data consists of the course name, tournament name, year, per round performance, finish position, as well as, many statistics rating the players performance at the given tournament. Below is an example of the format for the PGA Tour tournament history data set. The example consists of two players from four tournaments each with a sample of a few of the many important statistics. The statistics in the data set are, Fed Ex Points Won, Finish Position, Course Name (Long), Course Name (Short), Money Earned, Official Tournament, Perm Number, Player ID, Player Name, Round One Score, Round Two Score, Round Three Score, Round Four Score, Round Five Score, To Par Score, Total Par, Tournament Name and Year.

Player Name	Tournament Name	Course Name	Year	Round One Score	Finish Position
Patrick Reed	Arnold Palmer Invitational presented by Mastercard	Bay Hill Club & Lodge	2020	70	15
Patrick Reed	Arnold Palmer Invitational presented by Mastercard	Bay Hill Club & Lodge	2019	70	50
Patrick Reed	The Open Championship	Carnoustie GC	2018	75	28
Webb Simpson	Waste Management Phoenix Open	TPC Scottsdale	2020	71	1
Webb Simpson	Waste Management Phoenix Open	TPC Scottsdale	2019	67	20
Webb Simpson	RBC Canadian Open	St. George's G&CC	2010	70	37

2.5 Official World Golf Ranking (OWGR) Data

The data described above primarily takes into consideration players on the PGA Tour with an exception of a few statistics. In order to accurately make predictions when players from others tours compete on the PGA Tour, we must acquire data specific to other professional golf tours. The Official World Golf Ranking (OWGR) takes into consideration thirty-four professional golf tours. The number of players ranked by OWGR fluctuates, however there are typically approximately 9,000 athletes.

2.5.1 Example

The OWGR data set contains tournament data from 1985 to the most recent week. The data set is updated at the conclusion of each tournament. The data consists of the Player Name, Event, Tour, Year, Week, Finish, Rank Points, Weight, Adjusted Points, Rank After and Professional/Amateur. Below is an example of the format for the OWGR data set. The example consists of three players from two non PGA Tour tournaments each, with a sample of a few of the statistics.

Player Name	Tournament Name	Tour	Year	Week	Finish	Rank After
Sebastian Cappelen	Made In Denmark	EUR	2016	35	T26	488
Sebastian Cappelen	Midwest Classic	KFT	2014	30	T11	440
Wang Ter-Chang	Huangshan Open	CHN	2019	34	T20	2076
Wang Ter-Chang	Brunei Open	ASA	2008	34	T16	545
Craig Ainsley	The Motocaddy Masters	EPT	2017	32	T31	1966
Craig Ainsley	The "FORE" Business Championship	EPT	2017	34	WD	1950

2.6 Ladies Professional Golf Association (LPGA) Data

The Ladies Professional Golf Association (LPGA Tour) is the top professional league for females around the world. The LPGA Tour records several statistics every week to analyze the performance of each athlete. The statistics can be divided into the following sub-categories, Money, Driving, Short Game, Scoring, Total Played and Points. An example of the statistics include, Official Money, Driving Accuracy, Putting Average, Rolex Player of the Year, and many more. Every statistic from each of the sub-categories has been scraped from the LPGA website. Within each statistic there are variables that help to indicate the players performance, including the rank of the player in the given statistic for the current week and the value of the statistic of interest. The data is stored as the sum or average for each statistic for the year-to-date through.

2.6.1 Example

The LPGA Tour Statistics data set contains tournament data from 1980 to the most recent week. The data set is updated at the conclusion of each tournament. Below is an example of the format for the LPGA Tour Statistics data set. The example consists of four players and four statistics from four years.

Name	Year	Statistic	Variable	Value
Kay Cockerill	1993	Putting Average	Rank	115
Kay Cockerill	1993	Putting Average	Putts Average	30.68
Annika Sorenstam	2003	Average Driving Distance	Rank	1
Annika Sorenstam	2003	Average Driving Distance	Average Driving Distance	269.8
Lydia Ko	2015	Sand Saves	Rank	1
Lydia Ko	2015	Sand Saves	Percentage	59.09
Brooke M. Henderson	2019	Victories	Rank	3
Brooke M. Henderson	2019	Victories	Wins	2

3 Data to be Acquired

The following section highlights data that would be valuable to acquire.

3.1 European Tour Statistics

The European Tour is one of the largest professional golf tours. Many of the top athletes in the world play tournaments from both the PGA Tour and the European Tour. Since athletes occasionally play between both tours, there is currently a gap in statistics when athletes compete in European events. The OWGR data considers their performance and the strength of the field, but it does not offer performance statistics, such as, driving distance, putting average, etc. Acquiring the European Tour statistics would be valuable to analyze player performance by categories.