Data Analysis and Design

Brett Mele

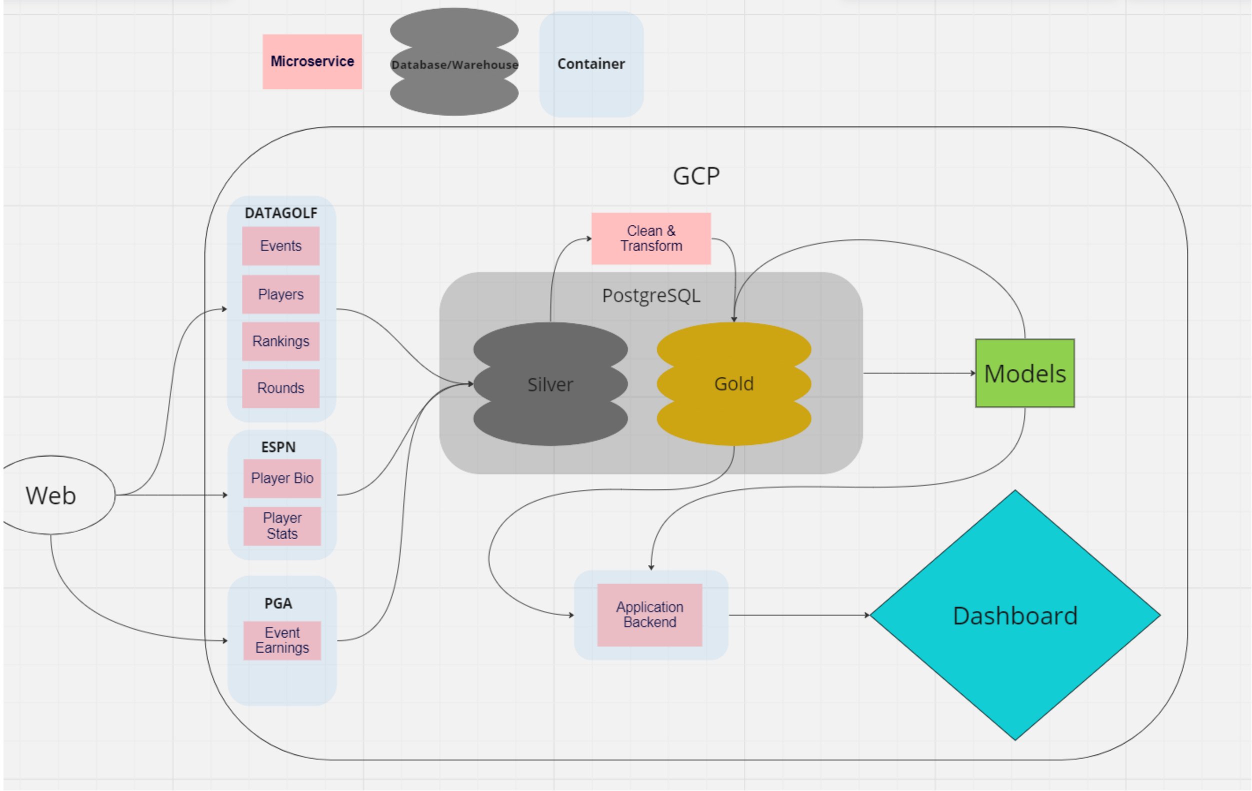
MSDS498 Northwestern University

July 16, 2023

Data Sources

Provider	Endpoint	Name	Description	Columns	Types
DataGolf	https://feeds.datagolf.com/historical-raw-data/event-list?file_format=json&key=	events	event list for all pro tour events	(calendar_year, date, event_id, event_name, has_sg, tour, has_traditional_stats)	(int, date, int, string, string, string, string)
DataGolf	https://feeds.datagolf.com/get-player-list?file_format=json&key=	players	player list for all players who played in a tour event since 2018	(dg_id, amateur, country_code, country, name)	(int, int, string, string, string)
DataGolf	https://feeds.datagolf.com/preds/get-dg-rankings?file_format=json&key=	rankings	datagolf model and official world golf rankings	(dg_id, player, primary_tour, amateur, country, dg_rank, dg_skill_estimate, owgr_rank, updated_at)	(int, string, string, int, string, int, float, int, datetime)
DataGolf	https://feeds.datagolf.com/historical-raw-data/rounds?tour=	rounds	round scores and stats for all PGA and Korn Ferry tour events since 2017	(tour, year, season, event_name, event_id, player_name, dg_id, fin_text, round_num, course_name, course_num,course_par, start_hole, teetime, round_scoure, sg_arg_sg_app, sg_off_tee, sg_t2g, sg_total, driving_dist, driving_acc, gir, scrambling, prox_rgh, prox_fw, great_shots, poor_shots)	(string, int, int, string, string, string, int, string, int, string, int, int, int, string, int, float, float, float, float, float, float, float, float, float, float, float, float, float, float)
ESPN.com	N/A	espn_bio	scraped player biographical information	(birthdate, birthplace, college, swing, turned_pro, link, espn_id)	(date, string, string, string, int, string, string)
ESPN.com	N/A	espn_stats	scraped player stats by year	(rk, name, age, earnings, cup, events, rnds, cuts, top10, wins, score, ddis, dacc, gir, putts, sand, birds, season, espn_id)	(int, string, int, float, int, int, int, int, int, int, float, float, float, float, float, float, float, int, string)
PGATour.com	N/A	earnings	scraped earnings by event dating back to 2015	(rank, player, money, tournament, season)	(int, string, int, string, string)

Data Management & Architecture



The project utilizes a cloud microservice architecture. The technology stack includes Google Cloud Platform (GCP), Postgres and Docker. Microservices for DataGolf are written in Go, while microservices that require web scraping are written in Python and rely on the BeautifulSoup and Selenium packages.

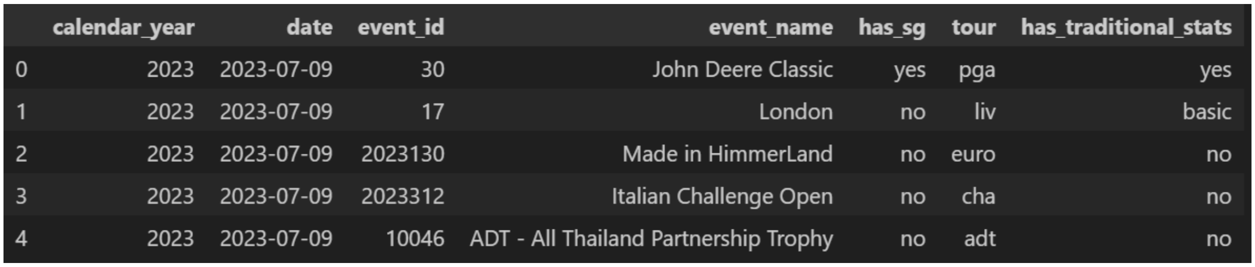
Data is sourced from the web, either from a REST API (using json format) or directly from the web page html. Once extracted, we write out tables to our "silver" Postgres database. Tables in this database are in their raw format directly from the API. Raw data from the silver tables is cleaned, transformed, merged and imported into our "gold" tables for use in analysis and modeling.

Merging of data sources will be required to meet our project objectives. DataGolf provides their own unique identifiers for events and players, which we will rely on for the majority of merges. Incorporating data from non-DataGolf sources will require creating reference tables containing mappings from each source to DataGolf player and event names. This will require a combination of automated matching based on player name (and tournament finish where applicable) and manual checks of those matches.

Data in gold tables will eventually need to be aggregated for use in models, but the exact nature of those aggregations is unclear until we begin model experimentation. We plan to determine the necessary aggregations via exploratory analysis and modeling, and create new gold tables where necessary to store aggregated information.

Analysis

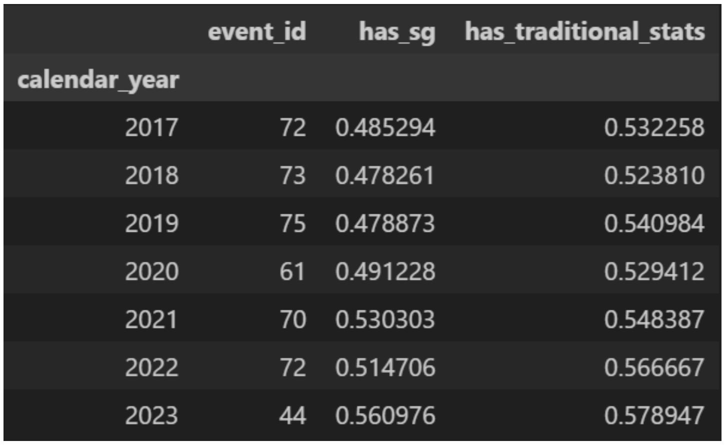
Events



	calendar_year	date	event_id	event_name	has_sg	tour	has_traditional_stats
0	2023	2023-07-09	30	John Deere Classic	yes	pga	yes
1	2023	2023-07-09	17	London	no	liv	basic
2	2023	2023-07-09	2023130	Made in HimmerLand	no	euro	no
3	2023	2023-07-09	2023312	Italian Challenge Open	no	cha	no
4	2023	2023-07-09	10046	ADT - All Thailand Partnership Trophy	no	adt	no

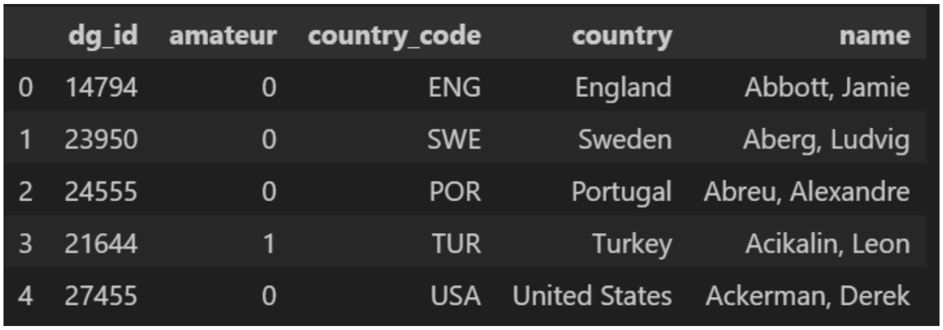
The event dataset contains the start date and other metadata regarding tour events dating back to 2017. This data joins to round scoring information on calendar_year and event_id.

One potential issue may be the number of events that have strokes gained and traditional stats. Only about 50% of events on the Korn Ferry and PGA tours have data within each year.



	event_id	has_sg	has_traditional_stats
calendar_year			
2017	72	0.485294	0.532258
2018	73	0.478261	0.523810
2019	75	0.478873	0.540984
2020	61	0.491228	0.529412
2021	70	0.530303	0.548387
2022	72	0.514706	0.566667
2023	44	0.560976	0.578947

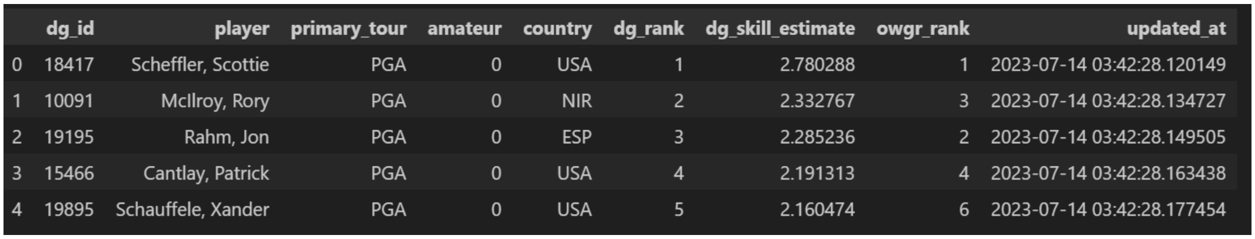
Players



	dg_id	amateur	country_code	country	name
0	14794	0	ENG	England	Abbott, Jamie
1	23950	0	SWE	Sweden	Aberg, Ludvig
2	24555	0	POR	Portugal	Abreu, Alexandre
3	21644	1	TUR	Turkey	Acikalin, Leon
4	27455	0	USA	United States	Ackerman, Derek

The players dataset contains DataGolf ids and biographical information on 3,281 players that played on all professional tours since 2017. This data joins to other tables with a dg_id foreign key.

Rankings

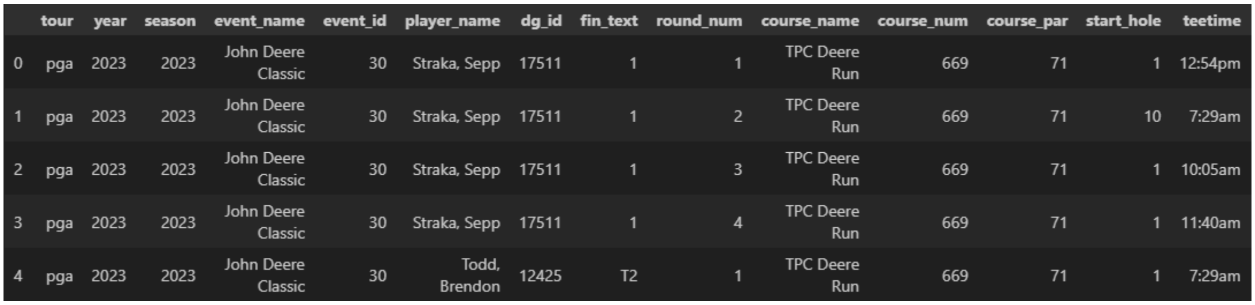


	dg_id	player	primary_tour	amateur	country	dg_rank	dg_skill_estimate	owgr_rank	updated_at
0	18417	Scheffler, Scottie	PGA	0	USA	1	2.780288	1	2023-07-14 03:42:28.120149
1	10091	McIlroy, Rory	PGA	0	NIR	2	2.332767	3	2023-07-14 03:42:28.134727
2	19195	Rahm, Jon	PGA	0	ESP	3	2.285236	2	2023-07-14 03:42:28.149505
3	15466	Cantlay, Patrick	PGA	0	USA	4	2.191313	4	2023-07-14 03:42:28.163438
4	19895	Schauffele, Xander	PGA	0	USA	5	2.160474	6	2023-07-14 03:42:28.177454

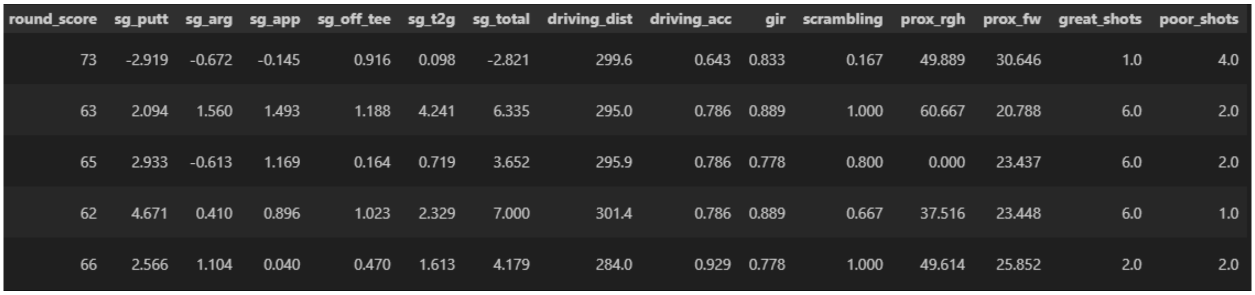
The rankings dataset contains player rankings for all players in the players dataset. Rankings include DataGolf proprietary model rankings and Official World Golf Rankings. Any player outside of the top 500 in either source is given a rank of 501.

Rankings data only includes the most recent rankings. No source for historical rankings data has been identified thus far. Thus any predictive models will likely not be able to rely on rankings as a feature. However, this data may be useful to present in the final dashboard.

Rounds



	tour	year	season	event_name	event_id	player_name	dg_id	fin_text	round_num	course_name	course_num	course_par	start_hole	teetime
0	pga	2023	2023	John Deere Classic	30	Straka, Sepp	17511	1	1	TPC Deere Run	669	71	1	12:54pm
1	pga	2023	2023	John Deere Classic	30	Straka, Sepp	17511	1	2	TPC Deere Run	669	71	10	7:29am
2	pga	2023	2023	John Deere Classic	30	Straka, Sepp	17511	1	3	TPC Deere Run	669	71	1	10:05am
3	pga	2023	2023	John Deere Classic	30	Straka, Sepp	17511	1	4	TPC Deere Run	669	71	1	11:40am
4	pga	2023	2023	John Deere Classic	30	Todd, Brendon	12425	T2	1	TPC Deere Run	669	71	1	7:29am

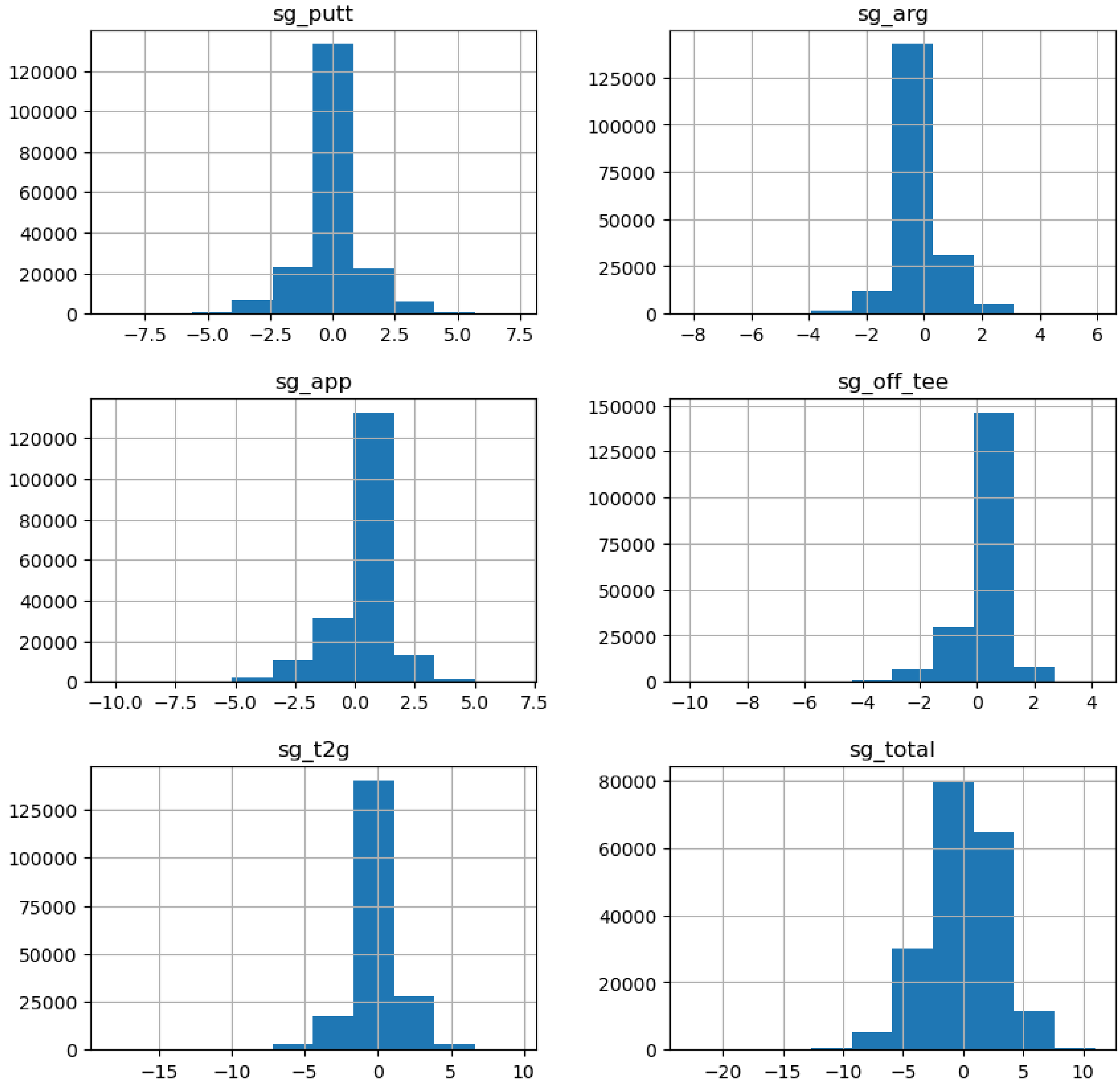


round_score	sg_putt	sg_arg	sg_app	sg_off_tee	sg_t2g	sg_total	driving_dist	driving_acc	gir	scrambling	prox_rgh	prox_fw	great_shots	poor_shots
73	-2.919	-0.672	-0.145	0.916	0.098	-2.821	299.6	0.643	0.833	0.167	49.889	30.646	1.0	4.0
63	2.094	1.560	1.493	1.188	4.241	6.335	295.0	0.786	0.889	1.000	60.667	20.788	6.0	2.0
65	2.933	-0.613	1.169	0.164	0.719	3.652	295.9	0.786	0.778	0.800	0.000	23.437	6.0	2.0
62	4.671	0.410	0.896	1.023	2.329	7.000	301.4	0.786	0.889	0.667	37.516	23.448	6.0	1.0
66	2.566	1.104	0.040	0.470	1.613	4.179	284.0	0.929	0.778	1.000	49.614	25.852	2.0	2.0

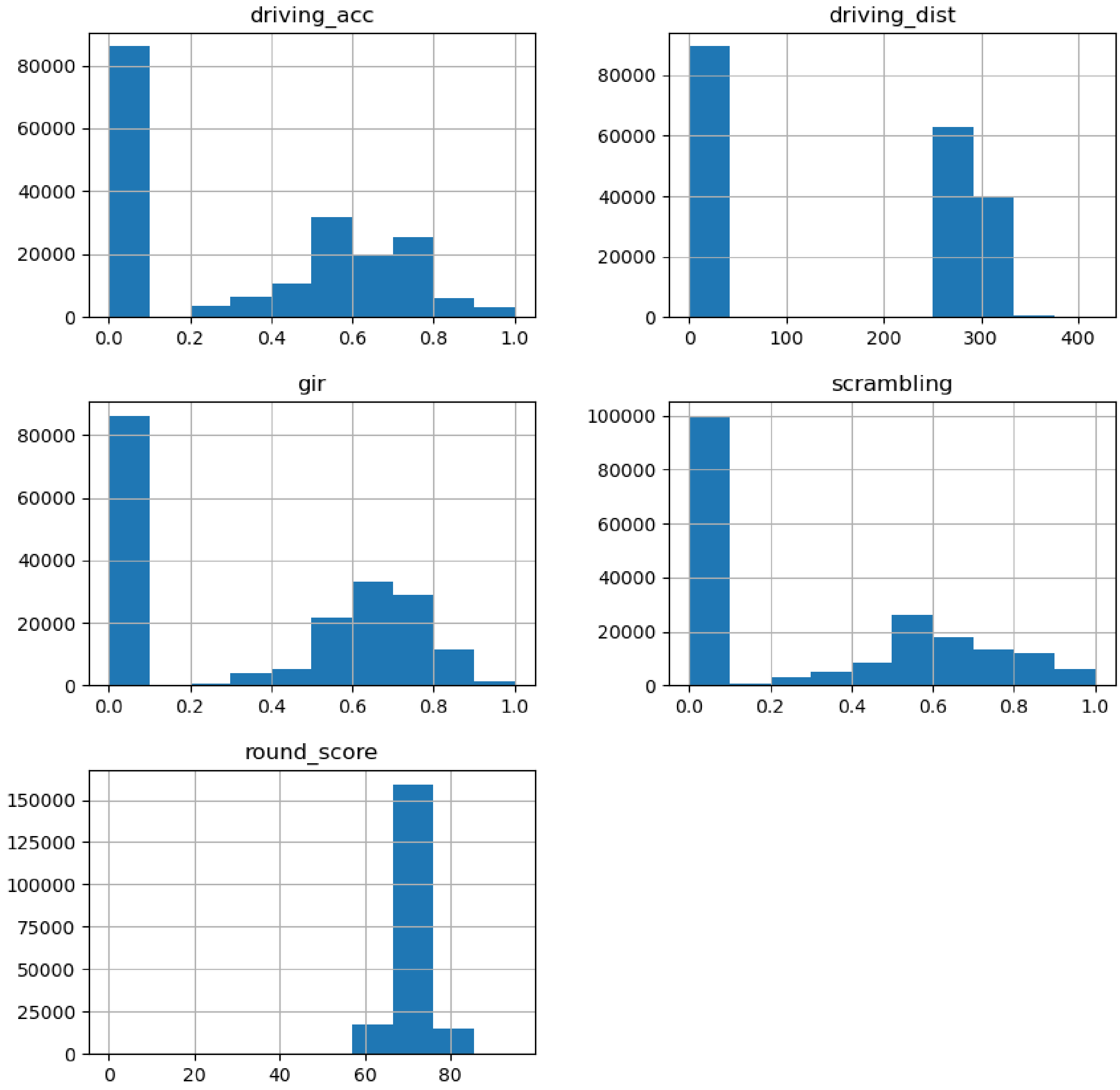
The rounds dataset will be the primary source used for predictive modeling. It contains strokes gained, traditional stats like driving distance and accuracy, as well as round and event metadata for all PGA and Korn Ferry events where data is available. This table contains foreign keys for event (year, event_id) and player (dg_id) metadata.

Preliminary exploratory analysis is shown below:

SG Distributions



Traditional stat distributions

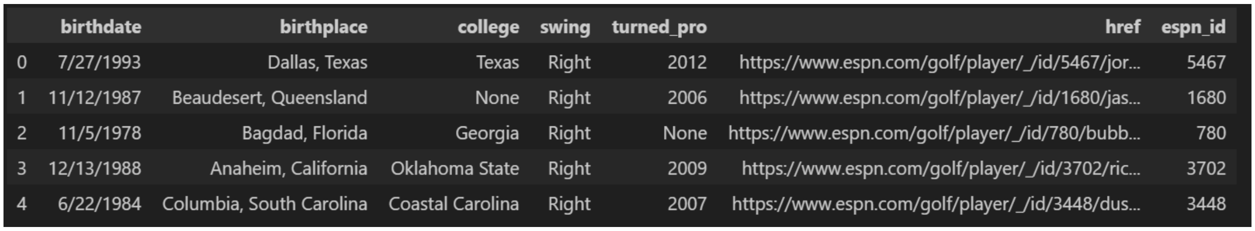


In the distributions, we see there is some zero inflation with most metrics. This is likely a result of missing data as Go enforces strict data types. We will handle these missing values in the data cleaning pipelines.

There are some important factors to consider in the round scoring dataset that will affect our modeling approach. First, only players who made the cut in a tournament will have data for all four rounds, so there is inherent selection bias that must be accounted for. We will also need to consider how to handle events where players dropped out due to injury, as well as how to handle the fact that players do not play every event within a season.

There is still more exploratory analysis to be done prior with this dataset prior to modeling. Once all data are cleaned we will conduct a more thorough investigation, including examining the relationships of target variables to features of interest.

ESPN Bio

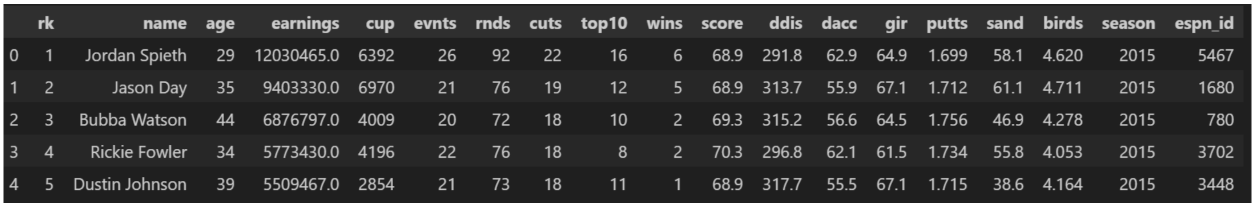


	birthdate	birthplace	college	swing	turned_pro	href	espn_id
0	7/27/1993	Dallas, Texas	Texas	Right	2012	https://www.espn.com/golf/player/_/id/5467/jor...	5467
1	11/12/1987	Beaudesert, Queensland	None	Right	2006	https://www.espn.com/golf/player/_/id/1680/jas...	1680
2	11/5/1978	Bagdad, Florida	Georgia	Right	None	https://www.espn.com/golf/player/_/id/780/bubb...	780
3	12/13/1988	Anaheim, California	Oklahoma State	Right	2009	https://www.espn.com/golf/player/_/id/3702/ric...	3702
4	6/22/1984	Columbia, South Carolina	Coastal Carolina	Right	2007	https://www.espn.com/golf/player/_/id/3448/dus...	3448

This dataset contains biographical data scraped from ESPN.com for all players who appeared on a professional tour since 2015. The primary use case for this dataset will be a source of player birth dates in order to get accurate ages at the start of each event.

There are two complicating issues with this dataset. First, not all players have a listed birthdate. For any such player, their age will have to be imputed. Second, there is no direct mapping between ESPN ids and DataGolf ids. We will have to curate a table that allows mapping between the two datasets via fuzzy name matching.

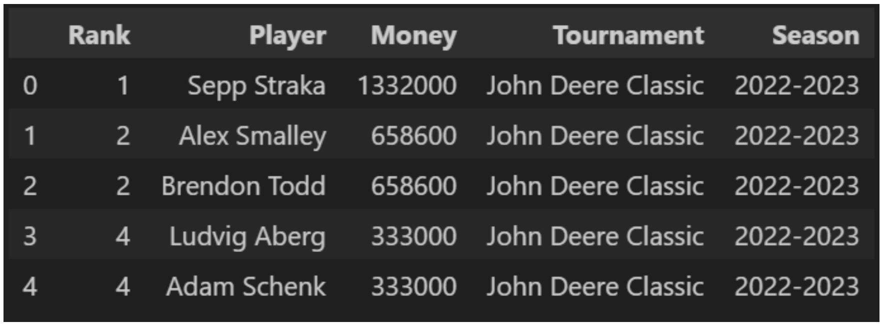
ESPN Stats



	rk	name	age	earnings	cup	evnts	rnds	cuts	top10	wins	score	ddis	dacc	gir	putts	sand	birds	season	espn_id
0	1	Jordan Spieth	29	12030465.0	6392	26	92	22	16	6	68.9	291.8	62.9	64.9	1.699	58.1	4.620	2015	5467
1	2	Jason Day	35	9403330.0	6970	21	76	19	12	5	68.9	313.7	55.9	67.1	1.712	61.1	4.711	2015	1680
2	3	Bubba Watson	44	6876797.0	4009	20	72	18	10	2	69.3	315.2	56.6	64.5	1.756	46.9	4.278	2015	780
3	4	Rickie Fowler	34	5773430.0	4196	22	76	18	8	2	70.3	296.8	62.1	61.5	1.734	55.8	4.053	2015	3702
4	5	Dustin Johnson	39	5509467.0	2854	21	73	18	11	1	68.9	317.7	55.5	67.1	1.715	38.6	4.164	2015	3448

This dataset contains yearly aggregate stats scraped from ESPN.com for players who appeared on the PGA tour in each season. It is currently unclear if this will provide any additional information not included in the DataGolf or earnings datasets, but may be useful to use as a cross reference when aggregating other sources. This table maps to espn_bio on espn_id.

Earnings



	Rank	Player	Money	Tournament	Season
0	1	Sepp Straka	1332000	John Deere Classic	2022-2023
1	2	Alex Smalley	658600	John Deere Classic	2022-2023
2	2	Brendon Todd	658600	John Deere Classic	2022-2023
3	4	Ludvig Aberg	333000	John Deere Classic	2022-2023
4	4	Adam Schenk	333000	John Deere Classic	2022-2023

This dataset scraped from PGATour.com contains tournament finish and earnings for each PGA tour event dating back to 2015. Earnings from this table will be used to model future earnings.

These data currently do not map directly to any of the other datasets, so similar to the espn data we will need to create a mapping to DataGolf player ids using name and tournament finish, and to DataGolf event ids using tournament name and season.

Notes

At this time we have collected all data sources and imported them into the silver Postgres database. There is still some work to be done to clean this data for use in exploratory analysis and modeling. We plan to create pipelines to clean these data and import into gold tables with proper data types and bad data removed. Part of this pipeline will include tables for source to source mappings of players and events.

We were able to obtain all of the necessary data sources to complete our project objectives. However, we have been unable to find data on course specific information and historical world golf rankings. Neither is imperative to meet our project goals, but both would help improve model performance if available.

One of the main challenges to address is the impact of incomplete data. Some events might not have associated stats, some players might not have birth dates available, and inherent in the structure of data there are issues of selection bias. It is not immediately clear the extent to which these issues will impact the final deliverables, as much of that will be addressed through further analysis and modeling.

Additionally, we were only able to collect data dating back to 2017. Since the goal of the project is to predict future performance and earnings, this may pose a challenge. The 2023 season is still ongoing, which leaves six seasons of complete data. This means the time horizon of our predictions will likely have to be reduced, or we will need to change our modeling approach (ex. make predictions for the following season and "chain" predictions together).