HRS2665 Applied Sports Science

Data hygiene

Data hygiene

What is it?

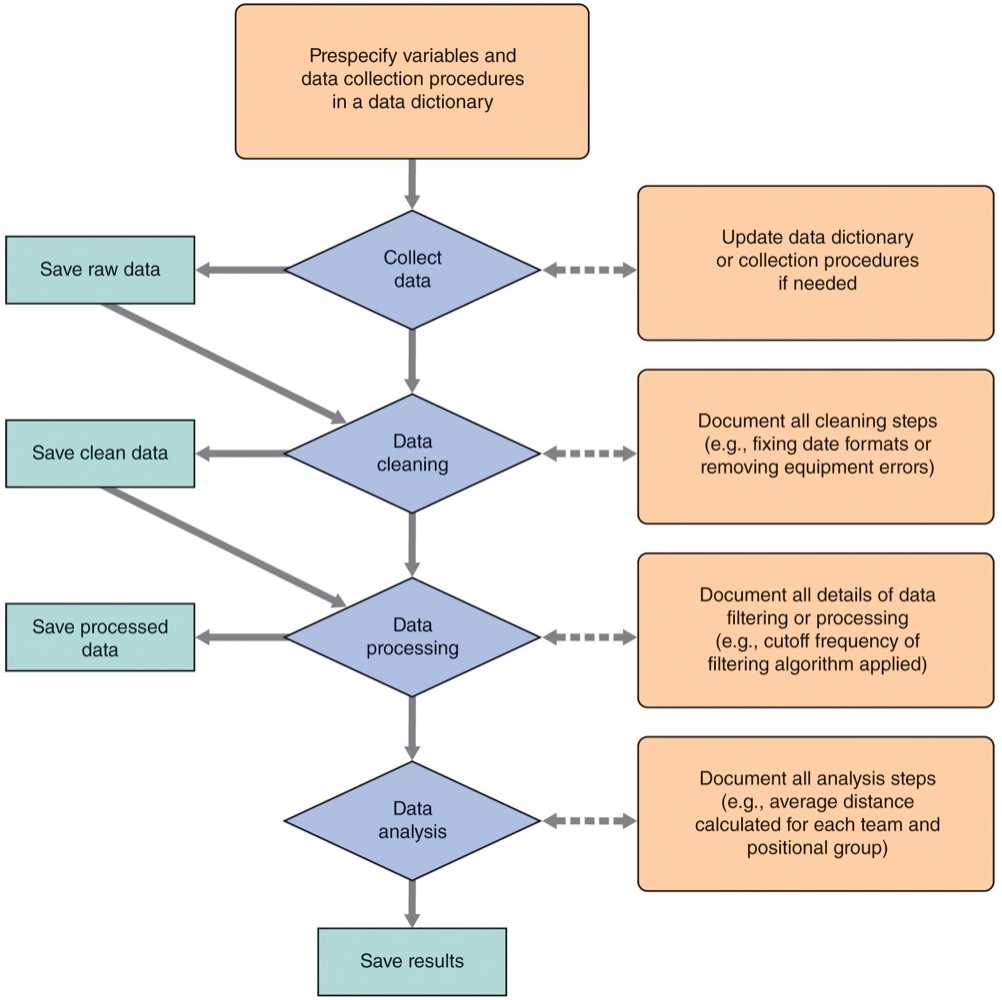
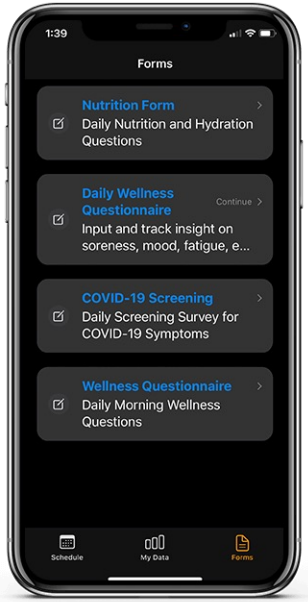


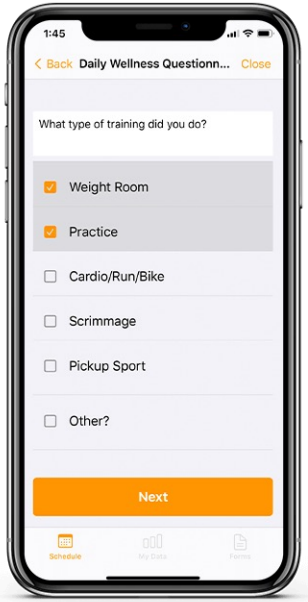
FIGURE 8.3 Documented and reproducible data workflow.

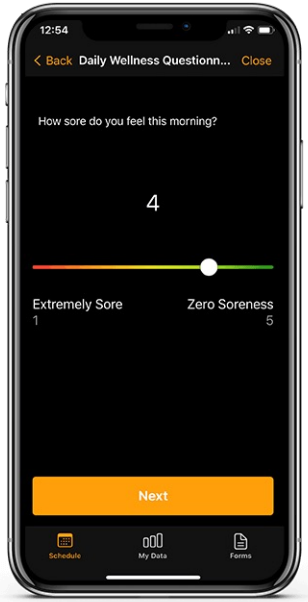
Data collection

Standardized nomenclature and acquisition parameters

Wellness Questionnaires







Standardization of:

- Question and assessment
  - Binary, categorical, continuous
- Answer modality
  - Slider, multiple choice, VAS, Yes/No
- Variable name
- Scale (if continuous)
- Special considerations:
  - Time/date
  - Open-ended questions / text box

Terminology & dictionary

Data collection

Nap	**Yes / No**
Nap_Duration	**Duration, increments of 1h**
Nap_duration	**Duration, increments of 0.5h**
Nap_quantity	**…**
Nap_quantity_h	**…**
Nap_time	**Time (HH:MM:SS) of start of nap**

ID	Date	Sport	Sex	Height	Weight
567943	04/03/2017	A	1	166	62
616852	22/05/2017	A	0	190	87
186451	22/05/2017	A	0	184	79
168321	14/11/2017	B	1	172	76

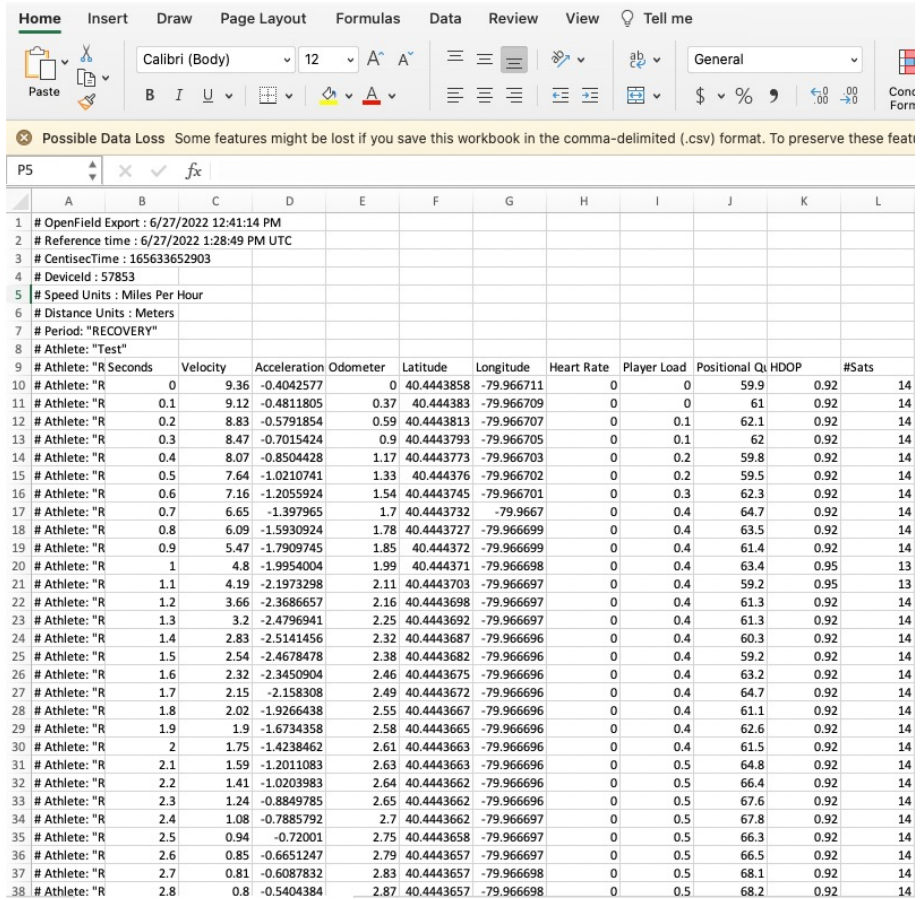
Variable	Definition
ID	Unique identification code (6-digit)
Date	Date that measurements were taken (Format: dd/mm/yyyy)
Sport	Sport played by participant (A: hockey; B: basketball)
Sex	Sex of participant (0: male; 1: female)
Height	Height of participant (cm) taken without shoes
Weight	Weight of participant (kg) taken without shoes

FIGURE 8.4 Sample data set and accompanying data dictionary.

Data structure

Raw vs. summary data

Raw



Home Insert Draw Page Layout Formulas Data Review View Tell me

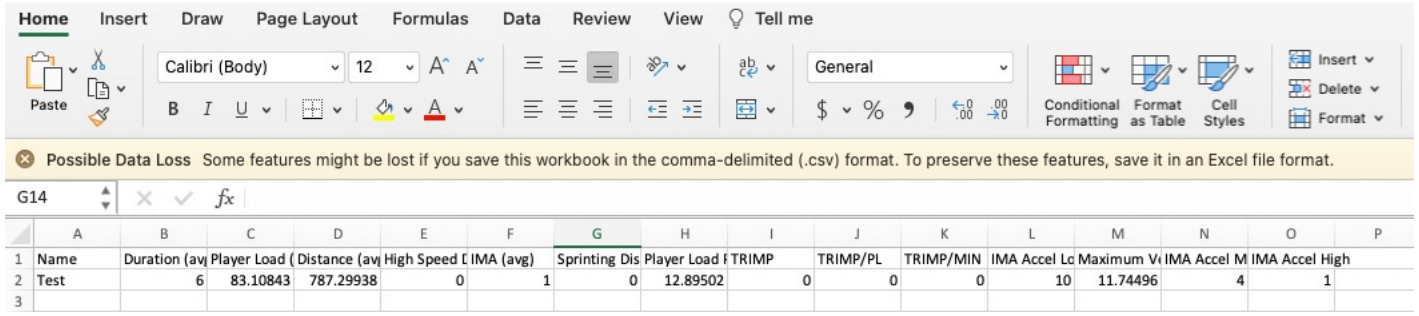
Paste | Calibri (Body) | 12 | General

Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these featu

P5

	A	B	C	D	E	F	G	H	I	J	K	L
1	# OpenField Export : 6/27/2022 12:41:14 PM											
2	# Reference time : 6/27/2022 1:28:49 PM UTC											
3	# CentisecTime : 165633652903											
4	# DeviceId : 57853											
5	# Speed Units : Miles Per Hour											
6	# Distance Units : Meters											
7	# Period: "RECOVERY"											
8	# Athlete: "Test"											
9	# Athlete: "R	Seconds	Velocity	Acceleration	Odometer	Latitude	Longitude	Heart Rate	Player Load	Positional Q	HDOP	#Sats
10	# Athlete: "R	0	9.36	-0.4042577	0	40.4443858	-79.966711	0	0	59.9	0.92	14
11	# Athlete: "R	0.1	9.12	-0.4811805	0.37	40.444383	-79.966709	0	0	61	0.92	14
12	# Athlete: "R	0.2	8.83	-0.5791854	0.59	40.4443813	-79.966707	0	0.1	62.1	0.92	14
13	# Athlete: "R	0.3	8.47	-0.7015424	0.9	40.4443793	-79.966705	0	0.1	62	0.92	14
14	# Athlete: "R	0.4	8.07	-0.8504428	1.17	40.4443773	-79.966703	0	0.2	59.8	0.92	14
15	# Athlete: "R	0.5	7.64	-1.0210741	1.33	40.444376	-79.966702	0	0.2	59.5	0.92	14
16	# Athlete: "R	0.6	7.16	-1.2055924	1.54	40.4443745	-79.966701	0	0.3	62.3	0.92	14
17	# Athlete: "R	0.7	6.65	-1.397965	1.7	40.4443732	-79.9667	0	0.4	64.7	0.92	14
18	# Athlete: "R	0.8	6.09	-1.5930924	1.78	40.4443727	-79.966699	0	0.4	63.5	0.92	14
19	# Athlete: "R	0.9	5.47	-1.7909745	1.85	40.444372	-79.966699	0	0.4	61.4	0.92	14
20	# Athlete: "R	1	4.8	-1.9954004	1.99	40.444371	-79.966698	0	0.4	63.4	0.95	13
21	# Athlete: "R	1.1	4.19	-2.1973298	2.11	40.4443703	-79.966697	0	0.4	59.2	0.95	13
22	# Athlete: "R	1.2	3.66	-2.3686657	2.16	40.4443698	-79.966697	0	0.4	61.3	0.92	14
23	# Athlete: "R	1.3	3.2	-2.4796941	2.25	40.4443692	-79.966697	0	0.4	61.3	0.92	14
24	# Athlete: "R	1.4	2.83	-2.5141456	2.32	40.4443687	-79.966696	0	0.4	60.3	0.92	14
25	# Athlete: "R	1.5	2.54	-2.4678478	2.38	40.4443682	-79.966696	0	0.4	59.2	0.92	14
26	# Athlete: "R	1.6	2.32	-2.3450904	2.46	40.4443675	-79.966696	0	0.4	63.2	0.92	14
27	# Athlete: "R	1.7	2.15	-2.158308	2.49	40.4443672	-79.966696	0	0.4	64.7	0.92	14
28	# Athlete: "R	1.8	2.02	-1.9266438	2.55	40.4443667	-79.966696	0	0.4	61.1	0.92	14
29	# Athlete: "R	1.9	1.9	-1.6734358	2.58	40.4443665	-79.966696	0	0.4	62.6	0.92	14
30	# Athlete: "R	2	1.75	-1.4238462	2.61	40.4443663	-79.966696	0	0.4	61.5	0.92	14
31	# Athlete: "R	2.1	1.59	-1.2011083	2.63	40.4443663	-79.966696	0	0.5	64.8	0.92	14
32	# Athlete: "R	2.2	1.41	-1.0203983	2.64	40.4443662	-79.966696	0	0.5	66.4	0.92	14
33	# Athlete: "R	2.3	1.24	-0.8849785	2.65	40.4443662	-79.966696	0	0.5	67.6	0.92	14
34	# Athlete: "R	2.4	1.08	-0.7885792	2.7	40.4443662	-79.966697	0	0.5	67.8	0.92	14
35	# Athlete: "R	2.5	0.94	-0.72001	2.75	40.4443658	-79.966697	0	0.5	66.3	0.92	14
36	# Athlete: "R	2.6	0.85	-0.6651247	2.79	40.4443657	-79.966697	0	0.5	66.5	0.92	14
37	# Athlete: "R	2.7	0.81	-0.6087832	2.83	40.4443657	-79.966698	0	0.5	68.1	0.92	14
38	# Athlete: "R	2.8	0.8	-0.5404384	2.87	40.4443657	-79.966698	0	0.5	68.2	0.92	14

Summary



Home Insert Draw Page Layout Formulas Data Review View Tell me

Paste | Calibri (Body) | 12 | General | Conditional Formatting | Format as Table | Cell Styles | Insert | Delete | Format

Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

G14

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Name	Duration (av	Player Load (	Distance (av	High Speed D	IMA (avg)	Sprinting Dis	Player Load F	TRIMP	TRIMP/PL	TRIMP/MIN	IMA Accel Lo	Maximum Ve	IMA Accel M	IMA Accel High	
2	Test	6	83.10843	787.29938	0	1	0	12.89502	0	0	0	10	11.74496	4	1	
3																

Many datapoints > 1 datapoint

Data cleaning

Tidy sheet

a

Sprint testing data

Player	Test 1 Date	Test 1 50 m sprint time (s)	Test 2 Date	Test 2 50 m sprint time (s)	Test 3 Date	Test 3 50 m sprint time (s)	Test 4 Date	Test 4 50 m sprint time (s)
J.Smith	17/02/2018	8.86	18/04/2018	7.29	18/08/2018	6.57	12/12/2018	6.86
S.Ford	14/03/2018	6.55	07/05/2018	injured	20/07/2018	6.44	06/03/2019	8.53
J.Brown	26/01/2018	6.79	16/04/2018	6.76	12/06/2018	7.76	23/03/2019	N/A
K.Adams	18/03/2018	7.60	12/05/2018	7.95	07/08/2018	missing	05/01/2019	6.38

b

	A	B	C	D	E	F
1	Player	ID_code	Date	Test_number	Time_50m_sec	Injured
2	J.Smith	P008	17/02/2018	1	8.86	0
3	S.Ford	P042	14/03/2018	1	6.55	0
4	J.Brown	P164	26/01/2018	1	6.79	0
5	K.Adams	P013	18/03/2018	1	7.60	0
6	J.Smith	P008	18/04/2018	2	7.29	0
7	S.Ford	P042	07/05/2018	2	NA	1
8	J.Brown	P164	16/04/2018	2	6.76	0
9	K.Adams	P013	12/05/2018	2	7.95	0
10	J.Smith	P008	18/08/2018	3	6.57	0
11	S.Ford	P042	20/07/2018	3	6.44	0
12	J.Brown	P164	12/06/2018	3	7.76	0
13	K.Adams	P013	07/08/2018	3	NA	0
14	J.Smith	P008	12/12/2018	4	6.86	0
15	S.Ford	P042	06/03/2019	4	8.53	0
16	J.Brown	P164	23/03/2019	4	NA	0
17	K.Adams	P013	05/01/2019	4	6.38	0

Strings, dates and factors

Conversion of variable types and units

Date

12/01/2022

01/12/2022

01/12/22

12-01-2022

12.01.2022

Dec-01-2022

01,December,2022

Thursday

String (character)

1 vs. “1”

Sleep_Duration vs. Sleep Duration

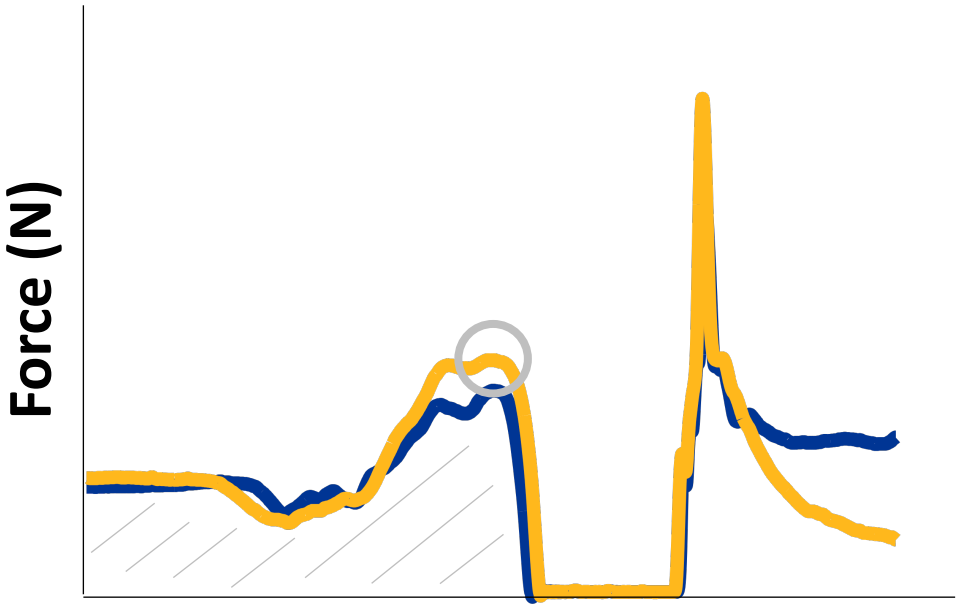
Factor

1, 2, 3, 4, 5 vs. one, two, three, four, five

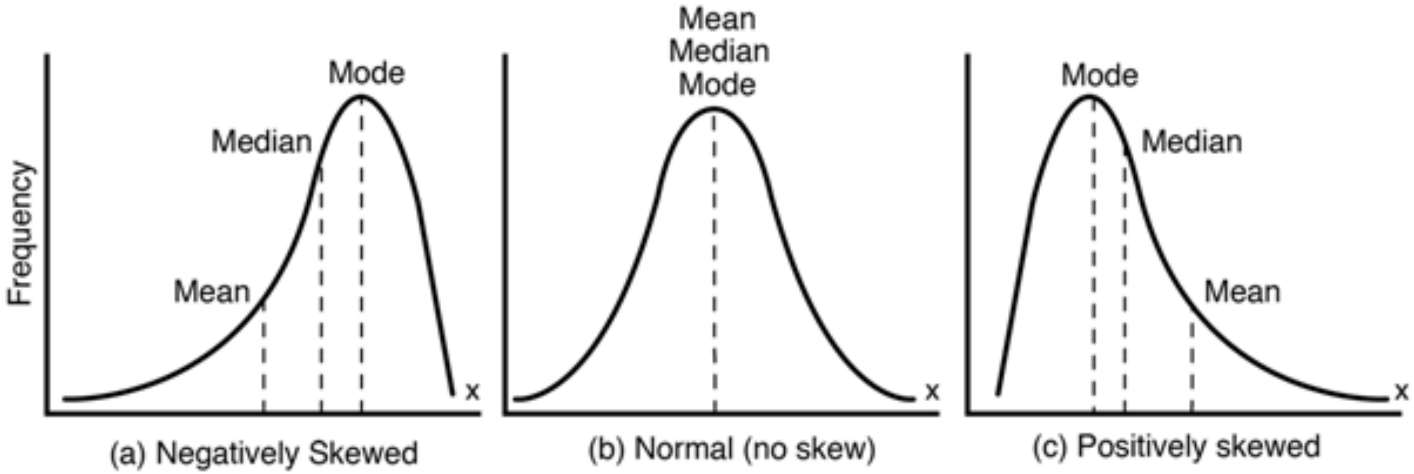
String (character) – Numeric – Factor – Boolean

Data processing

A world full of choices

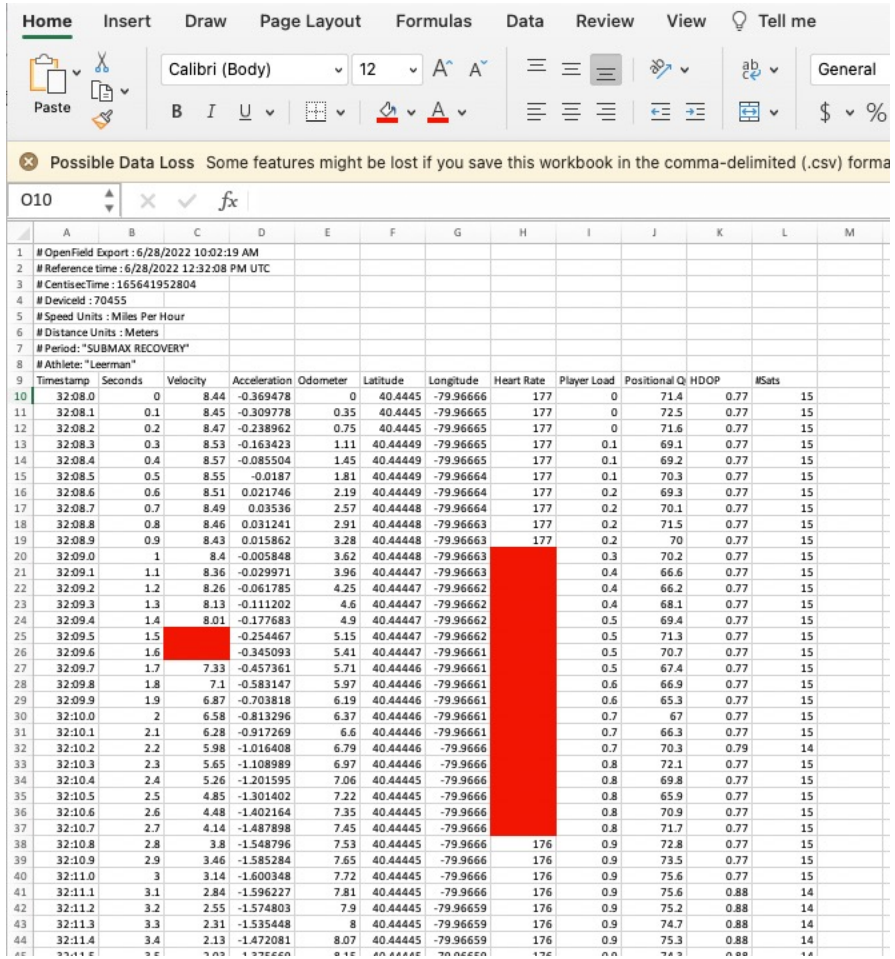


3 trials – which one do you analyze?



Data processing

Missing values, outlier detection



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	# OpenField Export : 6/28/2022 10:02:19 AM												
2	# Reference time : 6/28/2022 12:32:08 PM UTC												
3	# CentisecTime : 165641952804												
4	# DeviceId : 70455												
5	# Speed Units : Miles Per Hour												
6	# Distance Units : Meters												
7	# Period: "SUBMAX RECOVERY"												
8	# Athlete: "Leerman"												
9	Timestamp	Seconds	Velocity	Acceleration	Odometer	Latitude	Longitude	Heart Rate	Player Load	Positional Q	HDOP	#Sats	
10	32:08.0	0	8.44	-0.369478	0	40.4445	-79.96666	177	0	71.4	0.77	15	
11	32:08.1	0.1	8.45	-0.309778	0.35	40.4445	-79.96665	177	0	72.5	0.77	15	
12	32:08.2	0.2	8.47	-0.238962	0.75	40.4445	-79.96665	177	0	71.6	0.77	15	
13	32:08.3	0.3	8.53	-0.163423	1.11	40.44449	-79.96665	177	0.1	69.1	0.77	15	
14	32:08.4	0.4	8.57	-0.085504	1.45	40.44449	-79.96665	177	0.1	69.2	0.77	15	
15	32:08.5	0.5	8.55	-0.0187	1.81	40.44449	-79.96664	177	0.1	70.3	0.77	15	
16	32:08.6	0.6	8.51	0.021746	2.19	40.44449	-79.96664	177	0.2	69.3	0.77	15	
17	32:08.7	0.7	8.49	0.03536	2.57	40.44448	-79.96664	177	0.2	70.1	0.77	15	
18	32:08.8	0.8	8.46	0.031241	2.91	40.44448	-79.96663	177	0.2	71.5	0.77	15	
19	32:08.9	0.9	8.43	0.015862	3.28	40.44448	-79.96663	177	0.2	70	0.77	15	
20	32:09.0	1	8.4	-0.005848	3.62	40.44448	-79.96663		0.3	70.2	0.77	15	
21	32:09.1	1.1	8.36	-0.029971	3.96	40.44447	-79.96663		0.4	66.6	0.77	15	
22	32:09.2	1.2	8.26	-0.061785	4.25	40.44447	-79.96662		0.4	66.2	0.77	15	
23	32:09.3	1.3	8.13	-0.111202	4.6	40.44447	-79.96662		0.4	68.1	0.77	15	
24	32:09.4	1.4	8.01	-0.177683	4.9	40.44447	-79.96662		0.5	69.4	0.77	15	
25	32:09.5	1.5		-0.254467	5.15	40.44447	-79.96662		0.5	71.3	0.77	15	
26	32:09.6	1.6		-0.345093	5.41	40.44447	-79.96661		0.5	70.7	0.77	15	
27	32:09.7	1.7	7.33	-0.457361	5.71	40.44446	-79.96661		0.5	67.4	0.77	15	
28	32:09.8	1.8	7.1	-0.583147	5.97	40.44446	-79.96661		0.6	66.9	0.77	15	
29	32:09.9	1.9	6.87	-0.703818	6.19	40.44446	-79.96661		0.6	65.3	0.77	15	
30	32:10.0	2	6.58	-0.813296	6.37	40.44446	-79.96661		0.7	67	0.77	15	
31	32:10.1	2.1	6.28	-0.917269	6.6	40.44446	-79.96661		0.7	66.3	0.77	15	
32	32:10.2	2.2	5.98	-1.016408	6.79	40.44446	-79.9666		0.7	70.3	0.79	14	
33	32:10.3	2.3	5.65	-1.108989	6.97	40.44446	-79.9666		0.8	72.1	0.77	15	
34	32:10.4	2.4	5.26	-1.201595	7.06	40.44445	-79.9666		0.8	69.8	0.77	15	
35	32:10.5	2.5	4.85	-1.301402	7.22	40.44445	-79.9666		0.8	65.9	0.77	15	
36	32:10.6	2.6	4.48	-1.402164	7.35	40.44445	-79.9666		0.8	70.9	0.77	15	
37	32:10.7	2.7	4.14	-1.487898	7.45	40.44445	-79.9666		0.8	71.7	0.77	15	
38	32:10.8	2.8	3.8	-1.548796	7.53	40.44445	-79.9666	176	0.9	72.8	0.77	15	
39	32:10.9	2.9	3.46	-1.585284	7.65	40.44445	-79.9666	176	0.9	73.5	0.77	15	
40	32:11.0	3	3.14	-1.600348	7.72	40.44445	-79.9666	176	0.9	75.6	0.77	15	
41	32:11.1	3.1	2.84	-1.596227	7.81	40.44445	-79.9666	176	0.9	75.6	0.88	14	
42	32:11.2	3.2	2.55	-1.574803	7.9	40.44445	-79.96659	176	0.9	75.2	0.88	14	
43	32:11.3	3.3	2.31	-1.535448	8	40.44445	-79.96659	176	0.9	74.7	0.88	14	
44	32:11.4	3.4	2.13	-1.472081	8.07	40.44445	-79.96659	176	0.9	75.3	0.88	14	

When do we encounter missing values?

- Heart rate monitor
- Transfers / graduation
- Injuries / medical issue

When do we encounter outliers?

- Wrong calibration
- Catapult – HSD

Data visualization

Choice of graph type

Pitt SPORTS SCIENCE

Pittsburgh Panthe R s : : visualization

ggplot2.

The **ggplot2** package is built around layers that can be flexibly combined to create virtually any graph. Each graph is at least composed of **data, coordinates, geoms, facets and themes.**

— data
— coordinates
— geoms
— facets
— themes

ggplot2

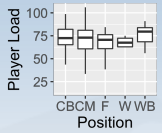
patchwork.

The **patchwork** package conveniently combines graphs to generate **layouts**, which can be shared with coaches. Each layout is a simple combination of ggplot2 figures.

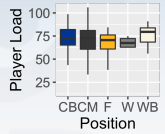
Horizontal: fig1 | fig2
Vertical: fig1 / fig2

Basics.

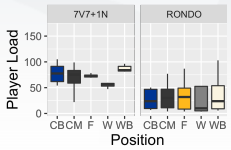
Define the dataset and coordinates.
fig <- df %>% ggplot(aes(x = Position, y = PL))



Define the type of graph via geoms.
fig + **geom_boxplot()**



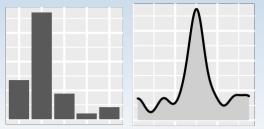
Define the color of factors via aesthetics.
fig + geom_boxplot(**aes(fill = Position)**)



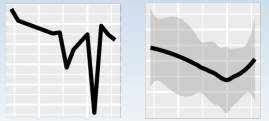
Visualize multiple factors via facets.
fig + geom_boxplot(aes(fill = position)) +
facet_wrap(~Drill)

Pitt Way.

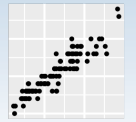
Cross-sectional
geom_bar() geom_density()



Time-series
geom_line() geom_smooth()

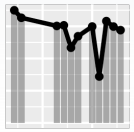


Correlation
geom_point()

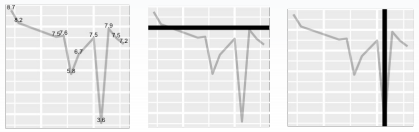


Profiling
geom_polygon() + coord_polar()

Mix & Match
geom_bar(alpha = 0.5) + geom_line() + geom_point()



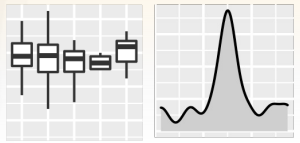
Contextualizing
geom_text() geom_hline() geom_vline()



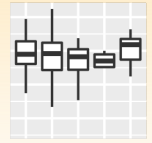
Basics.

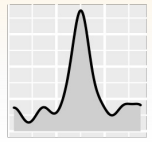
Create two figures.
fig1 <- fig + geom_boxplot()
fig2 <- fig + geom_density()

Combine horizontally.
fig1 | fig2



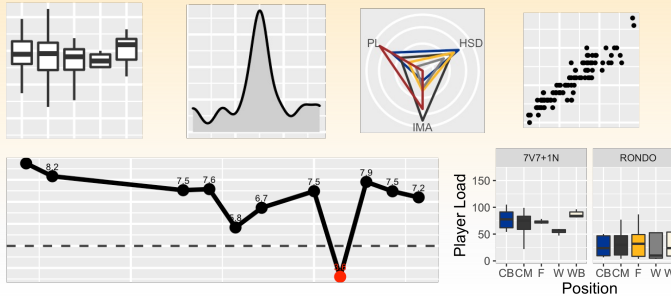
Combine vertically.
fig1 / fig2





Pitt Way.

Mix & Match and add spaces to generate daily reports.
(fig1 | fig2 | fig 3 | fig 4 | pitt_logo) / (fig6 | fig7 | **plot_spacer()**)

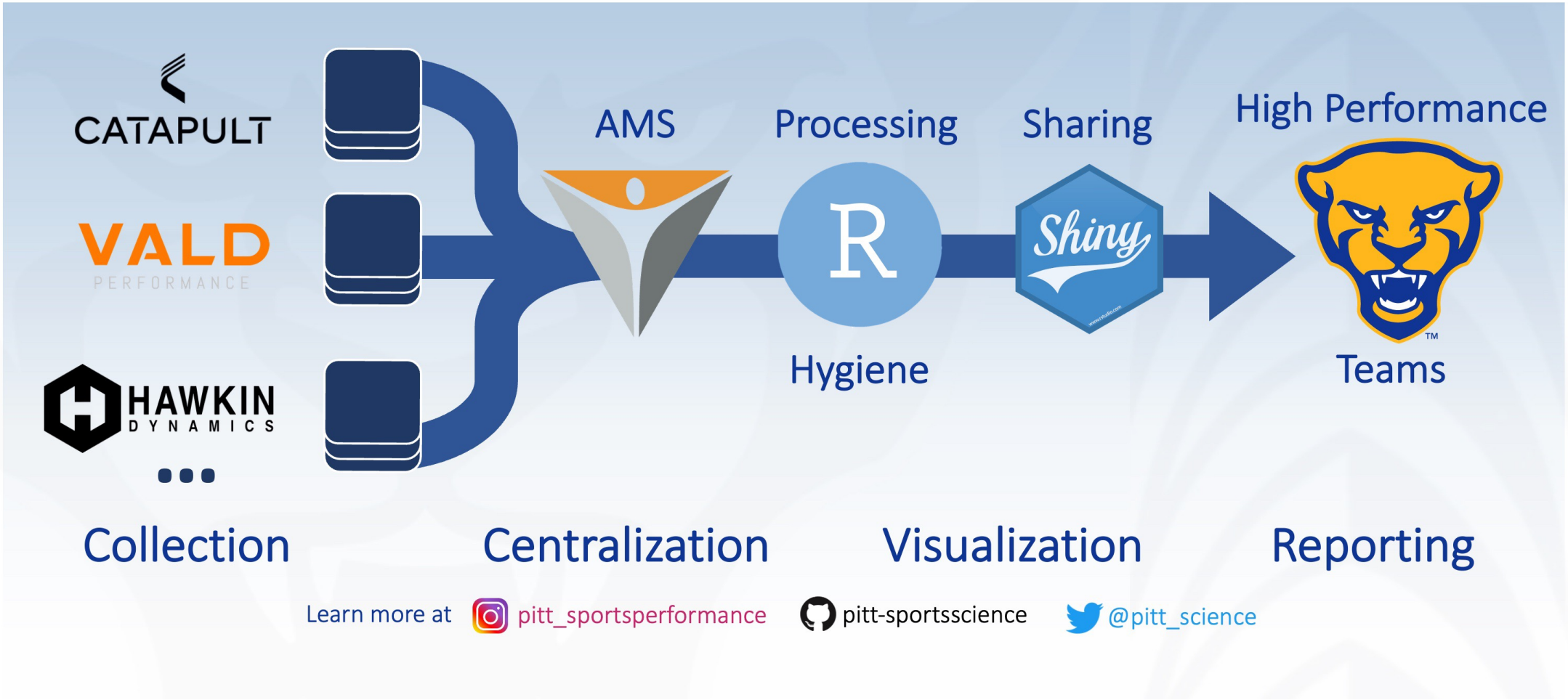


Learn more at
pitt_sportsperformance
pitt-sportsscience
@pitt_science

References: ggplot2: elegant graphics for data analysis. Hadley Wickham, Danielle Navarro, Thomas Lin Pedersen| https://ggplot2-book.org/ | https://patchwork.data-imaginist.com/ | @hadleywickham • @thomasp85 • @djnavarro | contact: fproessl@athletics.pitt.edu

Reproducible workflow

Coding languages



Questions?

Don’t forget to complete the quiz on Canvas