

# Playing with Data

Tobin Turner

2023-01-16



# Contents

<b>1</b>	<b>Reproducibility and Real Data</b>	<b>5</b>
1.1	Some Truth . . . . .	5
1.2	Critical Thinking, Analytics, and Reproducibility . . . . .	5
<b>2</b>	<b>Today's agenda</b>	<b>7</b>
<b>3</b>	<b>Marathon Kids</b>	<b>9</b>
3.1	About this data . . . . .	9
3.2	More about this data . . . . .	10
3.3	Some fun data for you . . . . .	10
3.4	The Hard Way . . . . .	12
<b>4</b>	<b>Starwars</b>	<b>13</b>
4.1	Data . . . . .	13
<b>5</b>	<b>NFL</b>	<b>15</b>
5.1	When do I go for it on 4th down? . . . . .	15
5.2	Data . . . . .	15
5.3	Reproducibility: Building is better . . . . .	15
<b>6</b>	<b>Some Final Thoughts</b>	<b>17</b>
6.1	Marathon Kids . . . . .	17
6.2	Starwars . . . . .	18



## Chapter 1

# Reproducibility and Real Data

### 1.1 Some Truth

“All models are wrong, but some are useful.”

– George Box, 1976, *Journal of the American Statistical Association*

### 1.2 Critical Thinking, Analytics, and Reproducibility



## Chapter 2

### Today's agenda

- Vaccines
- Stanford's President
- Target
- Dataset 1 (Marathon Kids; size, means and correlation?)
- Dataset 2 (Starwars, BMIs by homeworld?)
- Dataset 3 (NFL; 4th down?)
- See if we've had fun





## Chapter 3

# Marathon Kids

“Compare the means and correlations across trainers.”

### 3.1 About this data

trainer	pre	post
1	55.3846	97.1795
1	51.5385	96.0256
1	46.1538	94.4872
1	42.8205	91.4103
1	40.7692	88.3333
1	38.7179	84.8718

### 3.2 More about this data

trainer	n
1	142
2	142
3	142
4	142
5	142
6	142
7	142
8	142
9	142
10	142
11	142
12	142
13	142

### 3.3 Some fun data for you

Marathon Kids Data



### 3.4 The Hard Way

trainer	pre	post
1	55.38460	97.1795000
1	51.53850	96.0256000
1	46.15380	94.4872000
1	42.82050	91.4103000
1	40.76920	88.3333000
1	38.71790	84.8718000
1	35.64100	79.8718000
1	33.07690	77.5641000
1	28.97440	74.4872000
1	26.15380	71.4103000
1	23.07690	66.4103000
1	22.30770	61.7949000
1	22.30770	57.1795000
1	23.33330	52.9487000
1	25.89740	51.0256000
1	29.48720	51.0256000
1	32.82050	51.0256000
1	35.38460	51.4103000
1	40.25640	51.4103000
1	44.10260	52.9487000
1	46.66670	54.1026000
1	50.00000	55.2564000
1	53.07690	55.6410000
1	56.66670	56.0256000
1	59.23080	57.9487000
1	61.28210	62.1795000
1	61.53850	66.4103000
1	61.79490	69.1026000
1	57.43590	55.2564000
1	54.87180	49.8718000
1	52.56410	46.0256000
1	48.20510	38.3333000
1	49.48720	42.1795000
1	51.02560	44.1026000
1	45.38460	36.4103000
1	42.82050	32.5641000
1	38.71790	31.4103000
1	35.12820	30.2564000
1	32.56410	32.1795000
1	30.00000	36.7949000
1	33.58970	41.4103000
1	36.66670	45.6410000
1	38.20510	49.1026000
1	29.74360	36.0256000
1	29.74360	32.1795000
1	30.00000	29.1026000
1	32.05130	26.7949000
1	35.89740	25.2564000
1	41.02560	25.2564000
1	44.10260	25.6410000
1	47.17950	28.7180000

## Chapter 4

# Starwars

### 4.1 Data

Starwars Data

**Which homeworlds have the greatest number of individuals with BMI's greater than the average for each homework?**

name	height	mass	hair_color	skin_color	eye_color	birth
Luke Skywalker	172	77.0	blond	fair	blue	
C-3PO	167	75.0	NA	gold	yellow	
R2-D2	96	32.0	NA	white, blue	red	
Darth Vader	202	136.0	none	white	yellow	
Leia Organa	150	49.0	brown	light	brown	
Owen Lars	178	120.0	brown, grey	light	blue	
Beru Whitesun lars	165	75.0	brown	light	blue	
R5-D4	97	32.0	NA	white, red	red	
Biggs Darklighter	183	84.0	black	light	brown	
Obi-Wan Kenobi	182	77.0	auburn, white	fair	blue-gray	
Anakin Skywalker	188	84.0	blond	fair	blue	
Wilhuff Tarkin	180	NA	auburn, grey	fair	blue	
Chewbacca	228	112.0	brown	unknown	blue	
Han Solo	180	80.0	brown	fair	brown	
Greedo	173	74.0	NA	green	black	
Jabba Desilijic Tiure	175	1358.0	NA	green-tan, brown	orange	
Wedge Antilles	170	77.0	brown	fair	hazel	
Jek Tono Porkins	180	110.0	brown	fair	blue	
Yoda	66	17.0	white	green	brown	
Palpatine	170	75.0	grey	pale	yellow	
Boba Fett	183	78.2	black	fair	brown	
IG-88	200	140.0	none	metal	red	
Bossk	190	113.0	none	green	red	
Lando Calrissian	177	79.0	black	dark	brown	
Lobot	175	79.0	none	light	blue	
Ackbar	180	83.0	none	brown mottle	orange	
Mon Mothma	150	NA	auburn	fair	blue	
Arvel Crynyd	NA	NA	brown	fair	brown	
Wicket Systri Warrick	88	20.0	brown	brown	brown	
Nien Nunb	160	68.0	none	grey	black	
Qui-Gon Jinn	193	89.0	brown	fair	blue	
Nute Gunray	191	90.0	none	mottled green	red	
Finis Valorum	170	NA	blond	fair	blue	
Jar Jar Binks	196	66.0	none	orange	orange	
Roos Tarpals	224	82.0	none	grey	orange	
Rugor Nass	206	NA	none	green	orange	
Ric Olié	183	NA	brown	fair	blue	
Watto	137	NA	black	blue, grey	yellow	
Sebulba	112	40.0	none	grey, red	orange	
Quarsh Panaka	183	NA	black	dark	brown	
Shmi Skywalker	163	NA	black	fair	brown	
Darth Maul	175	80.0	none	red	yellow	
Bib Fortuna	180	NA	none	pale	pink	
Ayla Secura	178	55.0	none	blue	hazel	
Dud Bolt	94	45.0	none	blue, grey	yellow	
Gasgano	122	NA	none	white, blue	black	
Ben Quadinaros	163	65.0	none	grey, green, yellow	orange	
Mace Windu	188	84.0	none	dark	brown	
Ki-Adi-Mundi	198	82.0	white	pale	yellow	
Kit Fisto	196	87.0	none	green	black	
Eeth Koth	171	NA	black	brown	brown	
Adi Gallia	184	50.0	none	dark	blue	
Saesee Tiin	188	NA	none	pale	orange	

## Chapter 5

# NFL

### 5.1 When do I go for it on 4th down?

### 5.2 Data

NFL Data

NFL Descriptions

### 5.3 Reproducibility: Building is better

nflfastr





## Chapter 6

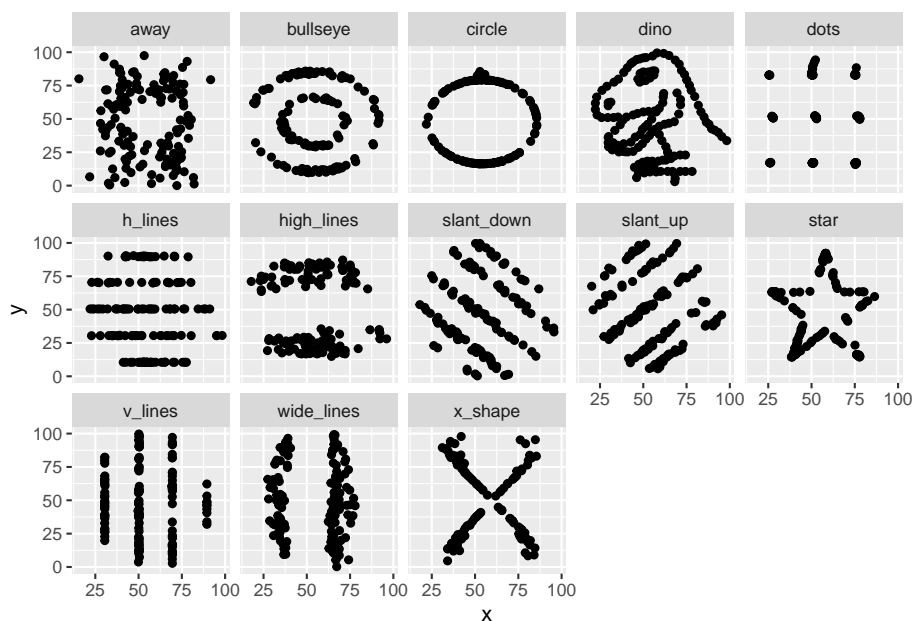
# Some Final Thoughts

### 6.1 Marathon Kids

#### 6.1.1 Mean and correlation Results

dataset	mean(x)	mean(y)	cor(x, y)
away	54.26610	47.83472	-0.0641284
bullseye	54.26873	47.83082	-0.0685864
circle	54.26732	47.83772	-0.0683434
dino	54.26327	47.83225	-0.0644719
dots	54.26030	47.83983	-0.0603414
h_lines	54.26144	47.83025	-0.0617148
high_lines	54.26881	47.83545	-0.0685042
slant_down	54.26785	47.83590	-0.0689797
slant_up	54.26588	47.83150	-0.0686092
star	54.26734	47.83955	-0.0629611
v_lines	54.26993	47.83699	-0.0694456
wide_lines	54.26692	47.83160	-0.0665752
x_shape	54.26015	47.83972	-0.0655833

### 6.1.2 A picture is worth a thousand words (or a dinosaur)



### 6.1.3 Reference

Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. Matejka, Fitzmaurice. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. May 2017. Pages 1290–1294. <https://doi.org/10.1145/3025453.3025912>.

## 6.2 Starwars

### 6.2.1 Missing values by variable

name	height	mass	homeworld	birth_year	species
0	6	28	10	44	4

### 6.2.2 BMI summary

name	height	mass	homeworld	BMI
Luke Skywalker	172	77	Tatooine	26.02758
C-3PO	167	75	Tatooine	26.89232
R2-D2	96	32	Naboo	34.72222
Darth Vader	202	136	Tatooine	33.33007
Leia Organa	150	49	Alderaan	21.77778
Owen Lars	178	120	Tatooine	37.87401

### 6.2.3 BMI summary

mean_bmi	median_bmi	max_bmi	min_bmi
32.01696	24.56749	443.4286	12.88625

### 6.2.4 Top contenders...

homeworld	count
Naboo	3
Tatooine	3
Alderaan	1
Corellia	1
Kamino	1
Kashyyyk	1
Mirial	1

### 6.2.5 And the winners are...

homeworld	count
Naboo	3
Tatooine	3

### 6.2.6 NFL, one option

Just one person's thoughts