Playing with Data

Tobin Turner

2023-01-21

Contents

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

Let's evaluate trainers. Calculate:

1. the `mean` pre for each trainer
2. the `mean` post for each trainer
3. the `correlation` between pre and post for each trainer

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 Full Data Set

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

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

4.1 Data

Starwars Data

4.2 Starwars missing values by variable

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

4.2.1 BMI summary

```
starwars %>% select(name, height, mass, homeworld) %>% na.omit() %>%
  mutate(BMI = mass/(height)^2*10000)
#> # A tibble: 56 x 5
#>     name               height  mass homeworld   BMI
#>     <chr>               <int> <dbl> <chr>     <dbl>
#>   1 Luke Skywalker        172    77 Tatooine   26.0
#>   2 C-3PO                 167    75 Tatooine   26.9
#>   3 R2-D2                  96    32 Naboo      34.7
#>   4 Darth Vader           202   136 Tatooine   33.3
#>  5 Leia Organa           150    49 Alderaan   21.8
#>  6 Owen Lars             178   120 Tatooine   37.9
#>  7 Beru Whitesun lars    165    75 Tatooine   27.5
#>  8 R5-D4                  97    32 Tatooine   34.0
#>  9 Biggs Darklighter     183    84 Tatooine   25.1
#> 10 Obi-Wan Kenobi        182    77 Stewjon    23.2
#> # ... with 46 more rows
```

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

4.3 BMI summary

mean_bmi	median_bmi	max_bmi	min_bmi
32.01696	24.56749	443.4286	12.88625

4.4 Top contenders...

```
starwars <- starwars %>%
  group_by(homeworld) %>%
  mutate(avg_bmi_by_hw = mean(BMI)) %>%
  ungroup()


above_avg_BMI <- starwars %>%
  filter(BMI > starwars$avg_bmi_by_hw) %>%
  group_by(homeworld) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
short_above <- head(above_avg_BMI, 7)
knitr::kable(short_above)
```

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

4.4.1 And the winners are...

homeworld	count
Naboo	3
Tatooine	3

Chapter 5

NFL

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

When?

5.2 Data

NFL Data

NFL Descriptions

5.3 Reproducibility: Building is better

nflfastr

Chapter 6

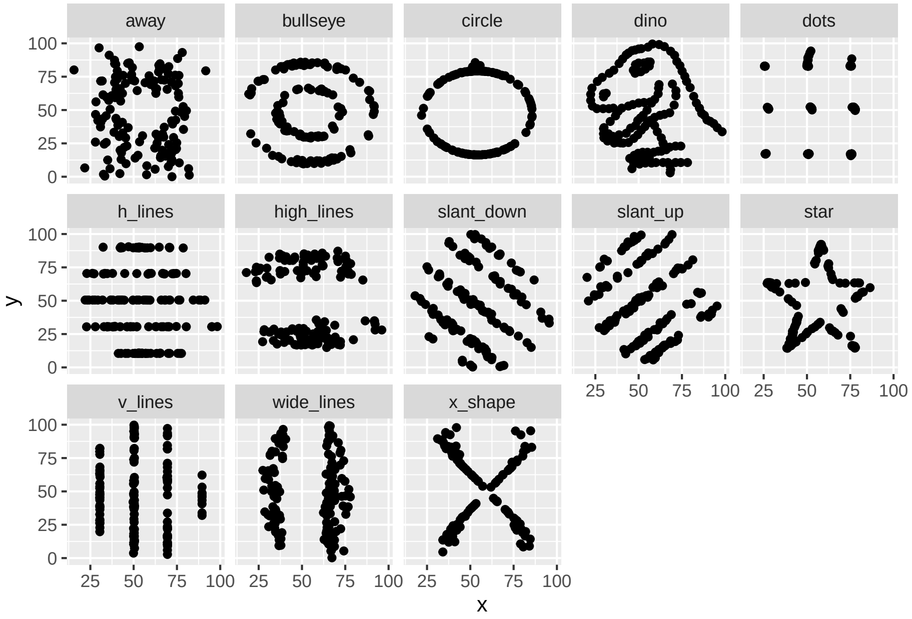
Some Final Thoughts

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.

6.1.4 NFL, one option

Just one person's thoughts