Playing with Data

Tobin Turner

2023-01-21

Contents

| 1 | Rep | producibility and Real Data | 5 |
|---|-----|---|-----------|
| | 1.1 | Some Truth | 5 |
| | 1.2 | Critical Thinking, Analytics, and Reproducibility | 5 |
| 2 | Tod | lay's agenda | 7 |
| 3 | Ma | rathon Kids | 9 |
| | 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 |
| 4 | Sta | rwars | 13 |
| | 4.1 | Data | 13 |
| 5 | NF | L | 15 |
| | 5.1 | When do I go for it on 4th down? | 15 |
| | 5.2 | Data | 15 |
| | 5.3 | Reproducibility: Building is better | 15 |
| 6 | Son | ne Final Thoughts | 17 |
| | 6.1 | Marathon Kids | 17 |
| | 6.0 | Ct | 10 |

4 CONTENTS

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

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

Marathon Kids

Let's evaluate traininers. Calculate: - the mean pre for each trainer - the mean post for each trainer - 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 Hard Way

| 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 22.30770 61.7949000 1 22.30770 57.1795000 1 22.30770 57.1795000 1 23.33330 52.9487000 1 22.30770 57.1795000 1 23.33330 52.9487000 1 23.33330 52.9487000 1 25.89740 51.0256000 1 32.82050 51.0256000 1 33.38460 51.4103000 1 44.10260 52.9487000 1 44.02660 52.9487000 1 50.0000 55.2564000 | | | |
|---|-------------|----------|------------|
| 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 23.07690 77.5641000 1 28.97440 74.4872000 1 26.15380 71.4103000 1 22.30770 61.7949000 1 22.30770 57.1795000 1 22.30770 57.1795000 1 22.38740 51.0256000 1 29.48720 51.0256000 1 32.82050 51.0256000 1 32.82050 51.0256000 1 35.38460 51.4103000 1 46.66670 54.1026000 1 46.66670 54.1026000 1 50.0000 55.2564000 1 50.0000 55.2564000 1 59.23080 57.9487000 1 | trainer | _ | post |
| 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 57.1795000 1 22.30770 57.1795000 1 22.30770 57.1795000 1 22.38740 51.0256000 1 29.48720 51.0256000 1 32.82050 51.0256000 1 32.82050 51.0256000 1 35.38460 51.4103000 1 44.10260 52.9487000 1 44.10260 52.9487000 1 45.66670 56.0256000 1 53.07690 55.6410000 1 55.236400 55.2564000 <t< th=""><th></th><th></th><th></th></t<> | | | |
| 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 57.1795000 1 22.30770 57.1795000 1 22.30770 57.1795000 1 22.39740 51.0256000 1 25.89740 51.0256000 1 32.82050 51.0256000 1 35.38460 51.4103000 1 44.10260 52.9487000 1 44.10260 52.9487000 1 44.10260 52.9487000 1 44.102600 55.2564000 1 50.0000 55.2564000 1 56.66670 56.0256000 1 57.43590 55.2564000 <td< th=""><th>1</th><th></th><th></th></td<> | 1 | | |
| 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 32.82050 51.0256000 1 35.38460 51.4103000 1 44.10260 52.9487000 1 44.10260 52.9487000 1 46.66670 54.1026000 1 53.07690 55.6410000 1 59.23080 57.9487000 1 61.28210 62.1795000 1 61.28210 62.1795000 1 61.53850 66.4103000 1 57.43590 55.2564000 <td< th=""><th>1</th><th></th><th></th></td<> | 1 | | |
| 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 35.38460 51.4103000 1 40.25640 51.4103000 1 44.10260 52.9487000 1 46.66670 54.1026000 1 50.0000 55.2564000 1 50.0000 55.6410000 1 50.0000 55.2564000 1 50.0000 55.2564000 1 50.0000 55.2564000 1 57.43590 55.2564000 1 57.43590 55.2564000 1 <th>1</th> <th></th> <th></th> | 1 | | |
| 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 57.1795000 1 22.30770 57.1795000 1 23.33330 52.9487000 1 25.89740 51.0256000 1 32.82050 51.0256000 1 35.38460 51.4103000 1 44.10260 52.9487000 1 44.10260 52.9487000 1 44.10260 52.9487000 1 46.66670 54.1026000 1 53.07690 55.6410000 1 53.07690 55.6410000 1 56.66670 56.0256000 1 59.23080 57.9487000 1 61.53850 66.4103000 1 61.53850 66.4103000 1 57.43590 55.2564000 <td< th=""><th>1</th><th></th><th></th></td<> | 1 | | |
| 1 33.07690 77.5641000 1 28.97440 74.4872000 1 26.15380 71.4103000 1 23.07690 66.4103000 1 22.30770 57.1795000 1 22.30770 57.1795000 1 23.33330 52.9487000 1 25.89740 51.0256000 1 32.82050 51.0256000 1 35.38460 51.4103000 1 44.10260 52.9487000 1 44.10260 52.9487000 1 44.10260 52.9487000 1 46.66670 54.1026000 1 50.00000 55.2564000 1 56.66670 56.0256000 1 59.23080 57.9487000 1 61.28210 62.1795000 1 61.79490 69.1026000 1 57.43590 55.2564000 1 52.56410 46.0256000 1 48.20510 38.3333000 <td< th=""><th>1</th><th></th><th></th></td<> | 1 | | |
| $\begin{array}{c ccccccccccccccccccccccccccccccccccc$ | 1 | | 79.8718000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | | 77.5641000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | 28.97440 | 74.4872000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | 26.15380 | 71.4103000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | 23.07690 | 66.4103000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | 22.30770 | 61.7949000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | 22.30770 | 57.1795000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | 23.33330 | 52.9487000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | 25.89740 | 51.0256000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | | 51.0256000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | 32.82050 | 51.0256000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | | 51.4103000 |
| $\begin{array}{c ccccccccccccccccccccccccccccccccccc$ | 1 | 40.25640 | |
| $\begin{array}{c ccccccccccccccccccccccccccccccccccc$ | 1 | 44.10260 | 52.9487000 |
| $\begin{array}{c ccccccccccccccccccccccccccccccccccc$ | 1 | 46.66670 | 54.1026000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | 50.00000 | 55.2564000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | | 53.07690 | 55.6410000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | 56.66670 | 56.0256000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | 59.23080 | 57.9487000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | 61.28210 | 62.1795000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | 61.53850 | 66.4103000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | 61.79490 | 69.1026000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | 57.43590 | 55.2564000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | 54.87180 | 49.8718000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | 52.56410 | 46.0256000 |
| $\begin{array}{c ccccccccccccccccccccccccccccccccccc$ | 1 | 48.20510 | 38.3333000 |
| $\begin{array}{c ccccccccccccccccccccccccccccccccccc$ | 1 | 49.48720 | 42.1795000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | 51.02560 | 44.1026000 |
| $\begin{array}{c ccccccccccccccccccccccccccccccccccc$ | 1 | 45.38460 | 36.4103000 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 1 | 42.82050 | 32.5641000 |
| $\begin{array}{c ccccccccccccccccccccccccccccccccccc$ | 1 | 38.71790 | 31.4103000 |
| $\begin{array}{c ccccccccccccccccccccccccccccccccccc$ | 1 | 35.12820 | 30.2564000 |
| $\begin{array}{c ccccccccccccccccccccccccccccccccccc$ | 1 | | |
| $\begin{array}{c ccccccccccccccccccccccccccccccccccc$ | 1 | 30.00000 | |
| $\begin{array}{c ccccccccccccccccccccccccccccccccccc$ | | | |
| $\begin{array}{c ccccccccccccccccccccccccccccccccccc$ | | | |
| $\begin{array}{c ccccccccccccccccccccccccccccccccccc$ | | | |
| $\begin{array}{c cccc} 1 & 29.74360 & 32.1795000 \\ \hline 1 & 30.00000 & 29.1026000 \\ \hline 1 & 32.05130 & 26.7949000 \\ \hline 1 & 35.89740 & 25.2564000 \\ \hline 1 & 41.02560 & 25.2564000 \\ \hline \end{array}$ | | | |
| 1 30.00000 29.1026000 1 32.05130 26.7949000 1 35.89740 25.2564000 1 41.02560 25.2564000 | | | |
| 1 32.05130 26.7949000 1 35.89740 25.2564000 1 41.02560 25.2564000 | | | |
| 1 35.89740 25.2564000 1 41.02560 25.2564000 | | | |
| 1 41.02560 25.2564000 | | | |
| | | | |
| 1 + 44.10260 + 25.6410000 | 1 | 44.10260 | 25.6410000 |
| 1 47.17950 28.7180000 | | | |

Starwars

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

4.1 Data

Starwars Data

NFL

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

NFL Data

 ${\it NFL Descriptions}$

5.3 Reproducibility: Building is better

nflfastr

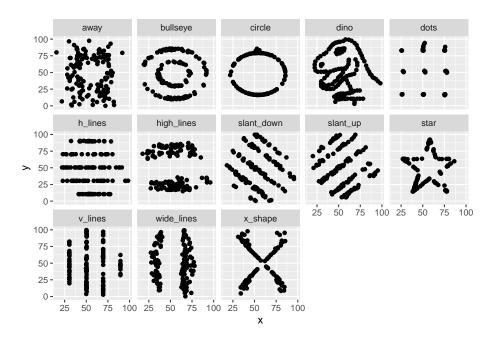
Some Final Thoughts

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. STARWARS 19

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