My title*

My subtitle if needed

First author Another author

November 28, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

*Code and data are available at: https://github.com/RohanAlexander/starter_folder.

2 Data

2.1 Measurement

2.2 Predictor Variables

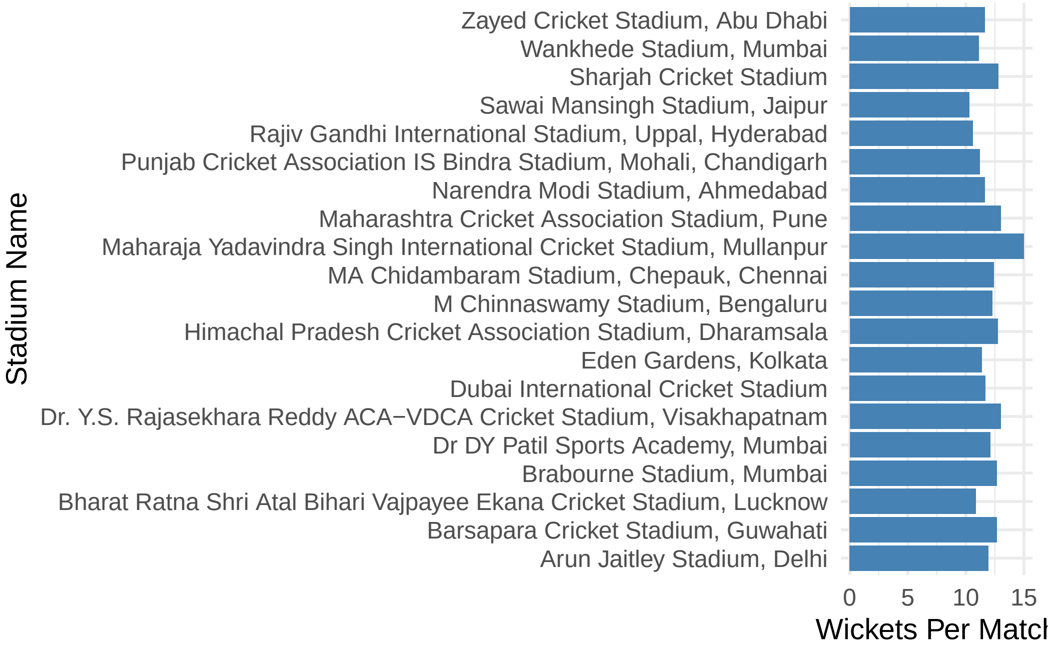
```r
num_bowlers_per_type <- cleaned_data %>%
  group_by(bowling_style) %>%
  summarise(
    num_bowlers = n_distinct(bowler)
  )

num_batters_per_type <- cleaned_data %>%
  group_by(batting_style) %>%
  summarise(
    num_bowlers = n_distinct(striker)
  )
```

2.3 Relationship Between Wickets and other Variables

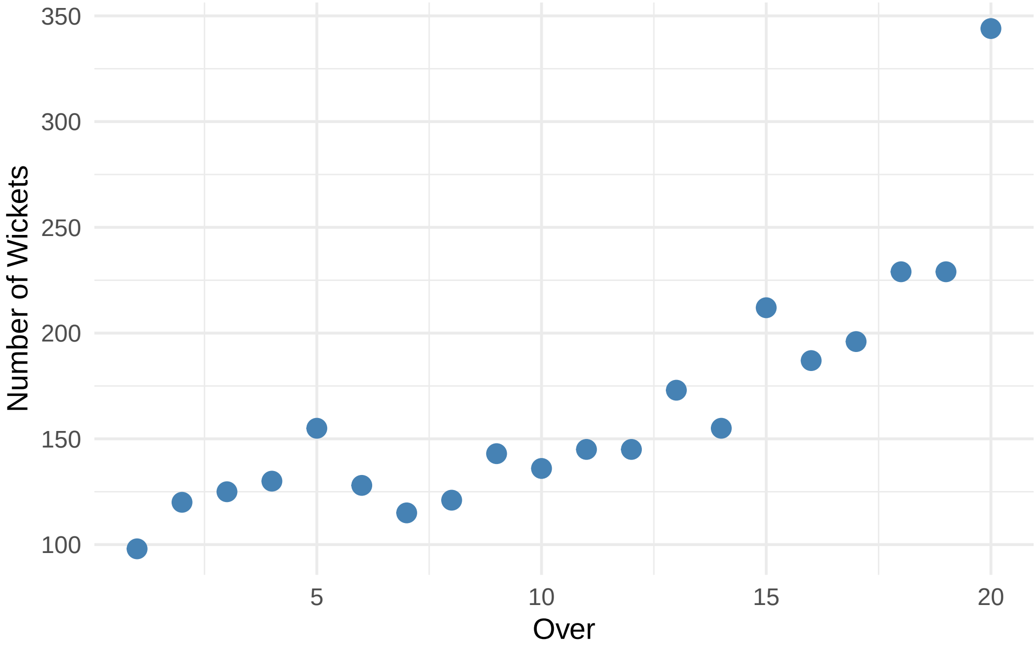
```r
cleaned_data %>% head()
```

```
# A tibble: 6 x 20
  match_id  year venue     innings  over  ball batting_team bowling_team striker
     <dbl> <dbl> <chr>       <dbl> <dbl> <dbl> <chr>        <chr>        <chr>
1  1254058  2021 MA Chida~       1     1     2 Mumbai Indi~ Royal Chall~ RG Sha~
2  1254058  2021 MA Chida~       1     1     3 Mumbai Indi~ Royal Chall~ RG Sha~
3  1254058  2021 MA Chida~       1     1     4 Mumbai Indi~ Royal Chall~ RG Sha~
4  1254058  2021 MA Chida~       1     1     5 Mumbai Indi~ Royal Chall~ RG Sha~
5  1254058  2021 MA Chida~       1     1     6 Mumbai Indi~ Royal Chall~ RG Sha~
6  1254058  2021 MA Chida~       1     2     1 Mumbai Indi~ Royal Chall~ RG Sha~
# i 11 more variables: bowler <chr>, runs_off_bat <dbl>,
#   wickets_lost_yet <dbl>, wicket <lgl>, target <dbl>, run_rate <dbl>,
#   batting_style <chr>, batter_playing_role <chr>, bowling_style <chr>,
#   bowler_playing_role <chr>, prev_over_wickets <int>
```

```
stadium_boundaries <- cleaned_data %>%
  group_by(venue) %>%
  summarise(
    num_matches = n_distinct(match_id),
    num_wickets = sum(wicket == TRUE),
  ) %>% arrange(desc(num_wickets), desc(num_matches))

ggplot(stadium_boundaries, aes(x = venue, y = (num_wickets/num_matches))) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(
       x = "Stadium Name",
       y = "Wickets Per Match") +
  theme_minimal() +
  coord_flip()
```



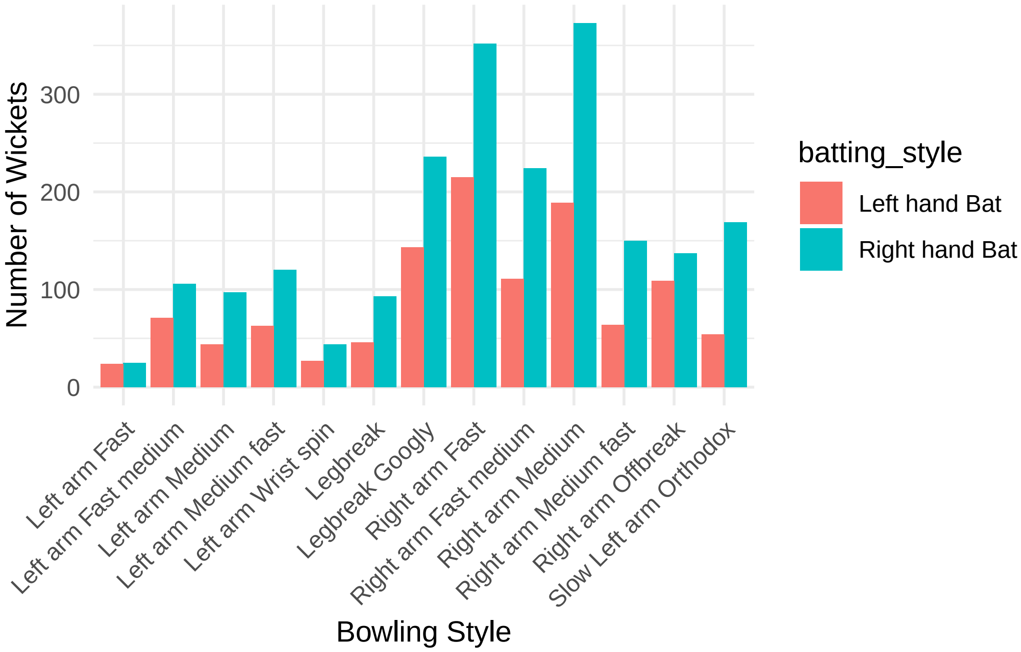
```
over_boundaries <- cleaned_data %>%
  group_by(over) %>%
  summarise(
    num_wickets = sum(wicket == TRUE),
    num_balls = n()
  ) %>% arrange(desc(num_wickets), desc(num_balls))

ggplot(over_boundaries, aes(x = over, y = num_wickets)) +
  geom_point(color = "steelblue", size = 3) +
  labs(
       x = "Over",
       y = "Number of Wickets") +
  theme_minimal()
```



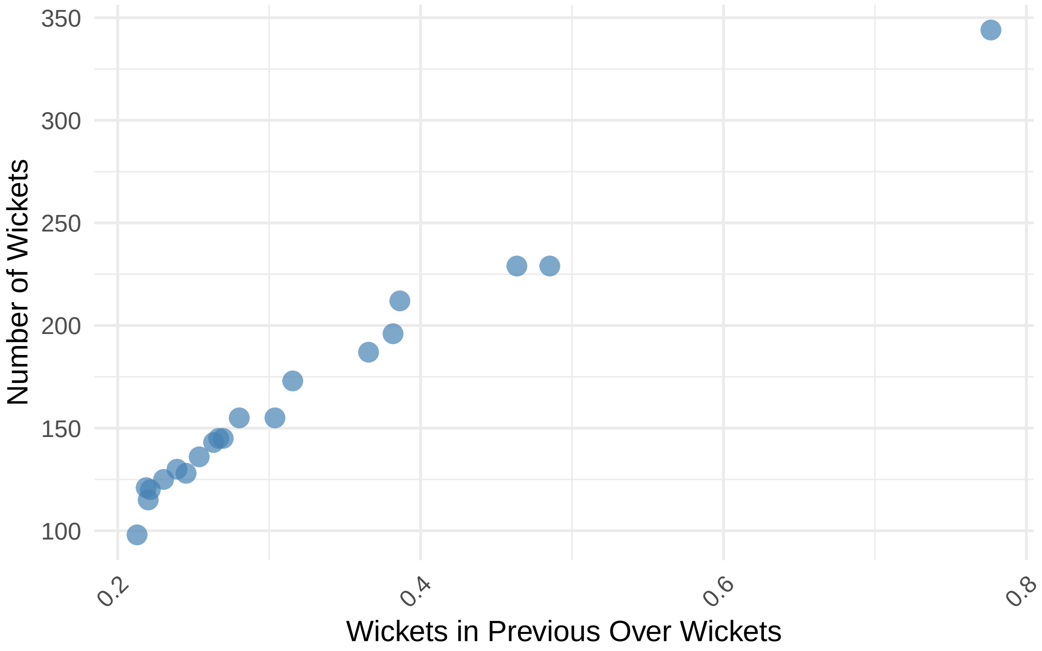
```
bowling_batting_matchup_boundaries <- cleaned_data %>%
  group_by(bowling_style, batting_style) %>%
  summarise(
    num_wickets = sum(wicket == TRUE),
    num_balls = n(),
  ) %>% arrange(desc(num_wickets), desc(num_balls))
```

```
`summarise()` has grouped output by 'bowling_style'. You can override using the
`.groups` argument.
```

```
ggplot(bowling_batting_matchup_boundaries, aes(x = bowling_style, y = num_wickets, fill = ba
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(
       x = "Bowling Style",
       y = "Number of Wickets") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
wickets_prev_over_wickets <- cleaned_data %>%
  group_by(over) %>%
  summarise(
    num_wickets = sum(wicket == TRUE),
    prev_over_wickets = mean(prev_over_wickets),
    num_balls = n(),
  ) %>% arrange(desc(num_wickets), desc(num_balls))

ggplot(wickets_prev_over_wickets, aes(x = prev_over_wickets, y = num_wickets)) +
  geom_point(color = "steelblue", size = 3, alpha = 0.7) +
  labs(
       x = "Wickets in Previous Over Wickets",
       y = "Number of Wickets") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



3 Model

3.1 Model set-up

3.1.1 Model justification

4 Results

```
simple_glm_wicket_model <-
  glm(
    wicket ~ over + wickets_lost_yet,
    data = cleaned_data,
    family = "binomial"
  )

summary(simple_glm_wicket_model)
```

```
Call:
glm(formula = wicket ~ over + wickets_lost_yet, family = "binomial",
    data = cleaned_data)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.399052   0.042537  -79.91   <2e-16 ***
over             -0.077155   0.005277  -14.62   <2e-16 ***
wickets_lost_yet  0.415210   0.011768   35.28   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25967  on 64508  degrees of freedom
Residual deviance: 24339  on 64506  degrees of freedom
AIC: 24345

Number of Fisher Scoring iterations: 6
```

```r
complex_glm_wicket_model <-
  glm(
    wicket ~ over + prev_over_wickets + wickets_lost_yet,
    data = cleaned_data,
    family = "binomial"
  )

summary(complex_glm_wicket_model)
```

```
Call:
glm(formula = wicket ~ over + prev_over_wickets + wickets_lost_yet,
    family = "binomial", data = cleaned_data)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -4.04743    0.04664  -86.79   <2e-16 ***
over              -0.06173    0.00534  -11.56   <2e-16 ***
prev_over_wickets  1.71746    0.02897   59.28   <2e-16 ***
wickets_lost_yet   0.23087    0.01320   17.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 25967  on 64508  degrees of freedom
Residual deviance: 20342  on 64505  degrees of freedom
AIC: 20350

Number of Fisher Scoring iterations: 6
```

```
overly_complex_glm_wicket_model <-
  glm(
    wicket ~ over + prev_over_wickets + batting_style + bowling_style,
    data = cleaned_data,
    family = "binomial"
  )

summary(overly_complex_glm_wicket_model)
```

```

Call:
glm(formula = wicket ~ over + prev_over_wickets + batting_style +
    bowling_style, family = "binomial", data = cleaned_data)

Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -4.372216   0.176157 -24.820   <2e-16 ***
over                                 0.005712   0.003474   1.644    0.100
prev_over_wickets                    1.854604   0.028232  65.691   <2e-16 ***
batting_styleRight hand Bat          0.044713   0.040721   1.098    0.272
bowling_styleLeft arm Fast medium    0.103461   0.190093   0.544    0.586
bowling_styleLeft arm Medium         0.266579   0.194117   1.373    0.170
bowling_styleLeft arm Medium fast    0.247459   0.188474   1.313    0.189
bowling_styleLeft arm Wrist spin     0.123592   0.216632   0.571    0.568
bowling_styleLegbreak                0.314879   0.193616   1.626    0.104
bowling_styleLegbreak Googly         0.248549   0.179581   1.384    0.166
bowling_styleRight arm Fast          0.227649   0.176652   1.289    0.198
bowling_styleRight arm Fast medium   0.211170   0.180708   1.169    0.243
bowling_styleRight arm Medium        0.194129   0.177249   1.095    0.273
bowling_styleRight arm Medium fast   0.221915   0.186790   1.188    0.235
bowling_styleRight arm Offbreak      0.114666   0.183713   0.624    0.533
bowling_styleSlow Left arm Orthodox  0.085553   0.185113   0.462    0.644
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25967  on 64508  degrees of freedom
Residual deviance: 20637  on 64493  degrees of freedom
AIC: 20669

Number of Fisher Scoring iterations: 6
```

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

```r
bowling_batting_role_matchup_boundaries <- cleaned_data %>%
  group_by(bowling_style, batter_playing_role) %>%
  summarise(
    num_wickets = sum(wicket == TRUE),
    num_balls = n(),
  ) %>% arrange(bowling_style, batter_playing_role)
```

```
`summarise()` has grouped output by 'bowling_style'. You can override using the
`.groups` argument.
```

B Model details

B.1 Posterior predictive check

C References

What to cite: - cricketdata - ESPNCricinfo - Cricsheet - kable - knitr - R - tidyverse - ggplot