My title[*]

My subtitle if needed

First author Another author

November 29, 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

```
cleaned_data %>% select(!c(ball, innings, match_id)) %>% head()
```

```
# A tibble: 6 x 17
   year venue          over batting_team bowling_team striker bowler runs_off_bat
  <dbl> <chr>         <dbl> <chr>        <chr>        <chr>   <chr>         <dbl>
1  2021 MA Chidamba~      1 Mumbai Indi~ Royal Chall~ RG Sha~ Moham~            0
2  2021 MA Chidamba~      1 Mumbai Indi~ Royal Chall~ RG Sha~ Moham~            0
3  2021 MA Chidamba~      1 Mumbai Indi~ Royal Chall~ RG Sha~ Moham~            2
4  2021 MA Chidamba~      1 Mumbai Indi~ Royal Chall~ RG Sha~ Moham~            0
5  2021 MA Chidamba~      1 Mumbai Indi~ Royal Chall~ RG Sha~ Moham~            1
6  2021 MA Chidamba~      2 Mumbai Indi~ Royal Chall~ RG Sha~ KA Ja~            1
# i 9 more variables: 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>
```

2.1 Measurement

2.2 Predictor Variables

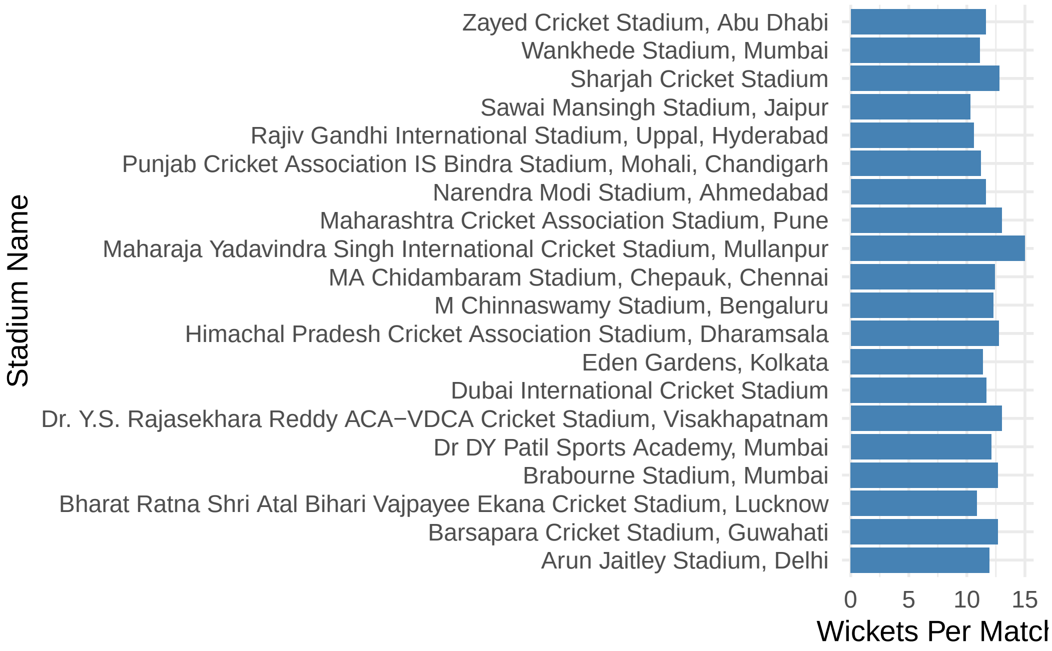
```
num_bowlers_per_type <- cleaned_data %>%
  group_by(bowling_style) %>%
  summarise(
    num_bowlers = n_distinct(bowler)
  )

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

2.3 Relationship Between Wickets and other Variables

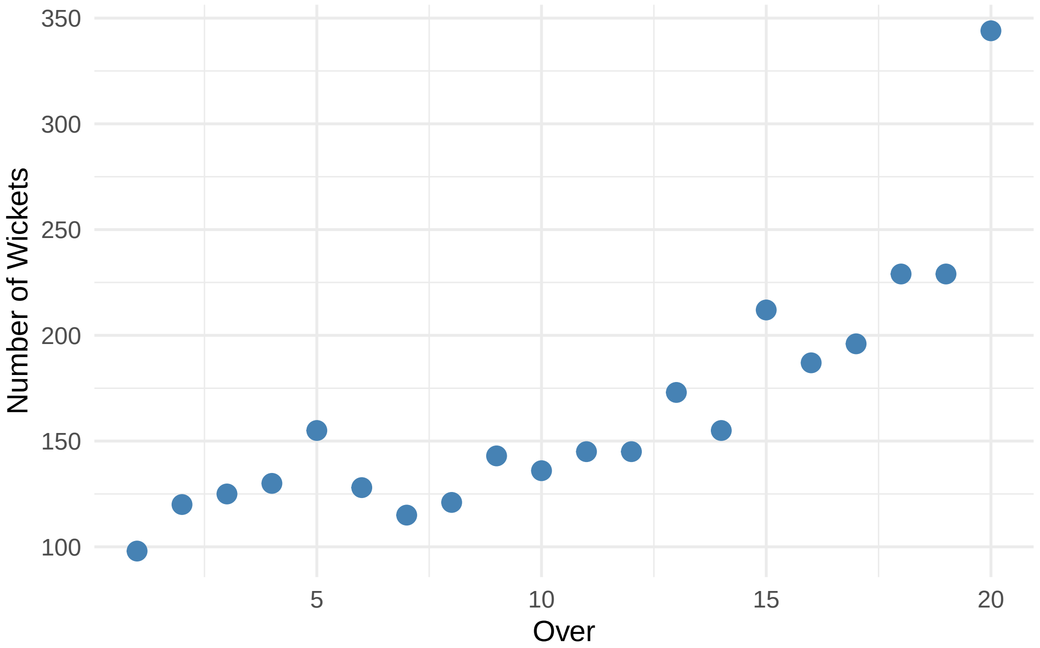
```
cleaned_data %>% head()
```

```
# 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>
```

```r
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()
```



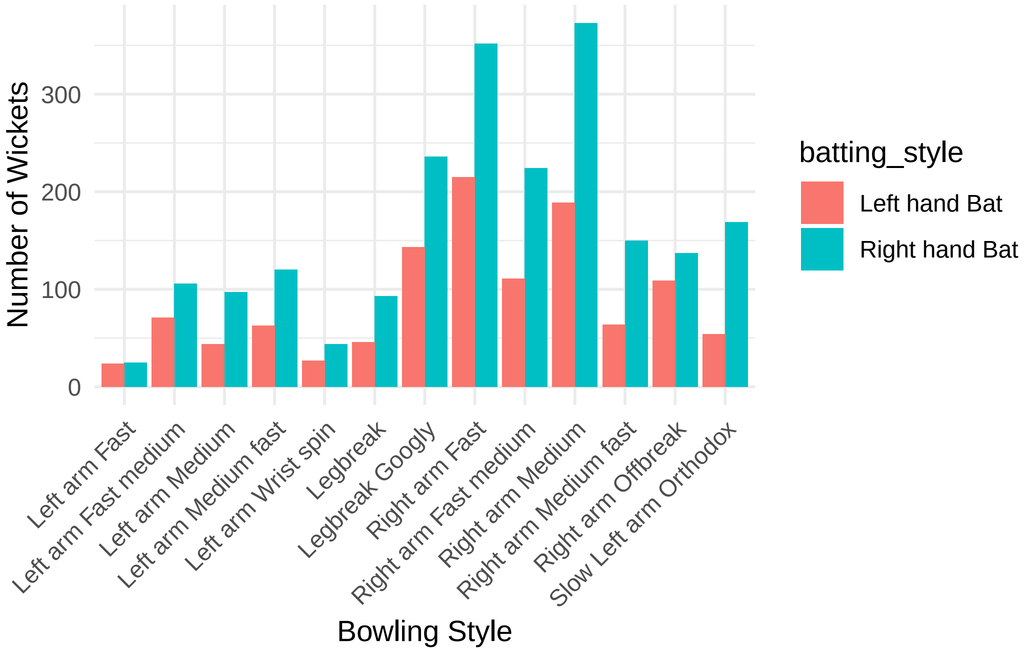
```r
over_boundaries <- cleaned_data %>%
  group_by(over) %>%
  summarise(
    num_wickets = sum(wicket == TRUE),
    num_balls = n()
  ) %>% arrange(desc(num_wickets), desc(num_balls))

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



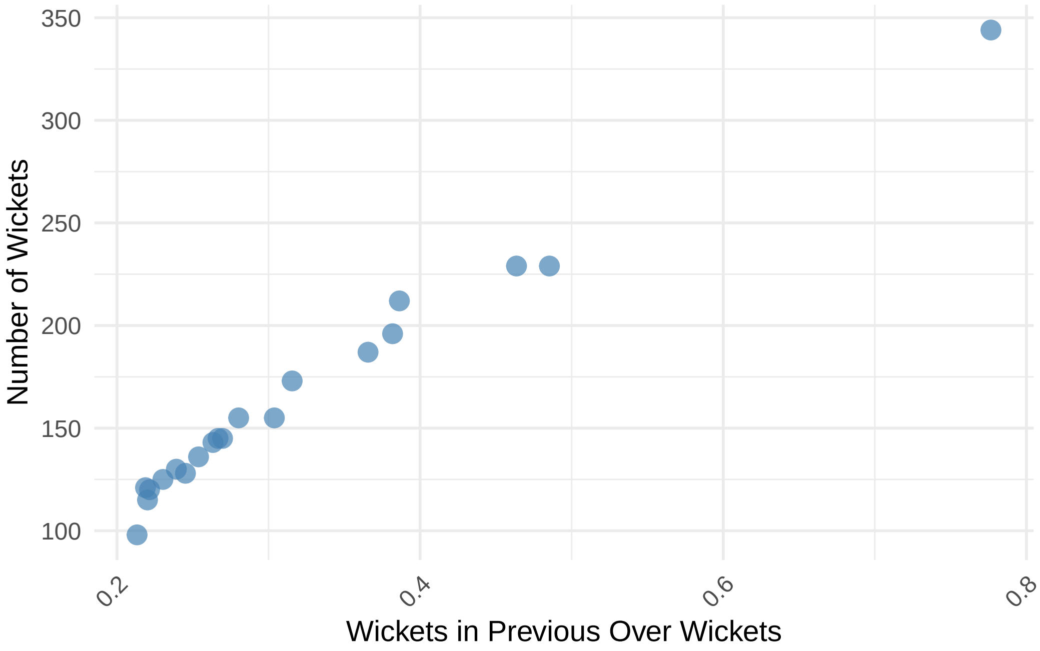
```
bowling_batting_matchup_boundaries <- cleaned_data %>%
  group_by(bowling_style, batting_style) %>%
  summarise(
    num_wickets = sum(wicket == TRUE),
    num_balls = n(),
  ) %>% arrange(desc(num_wickets), desc(num_balls))
```

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

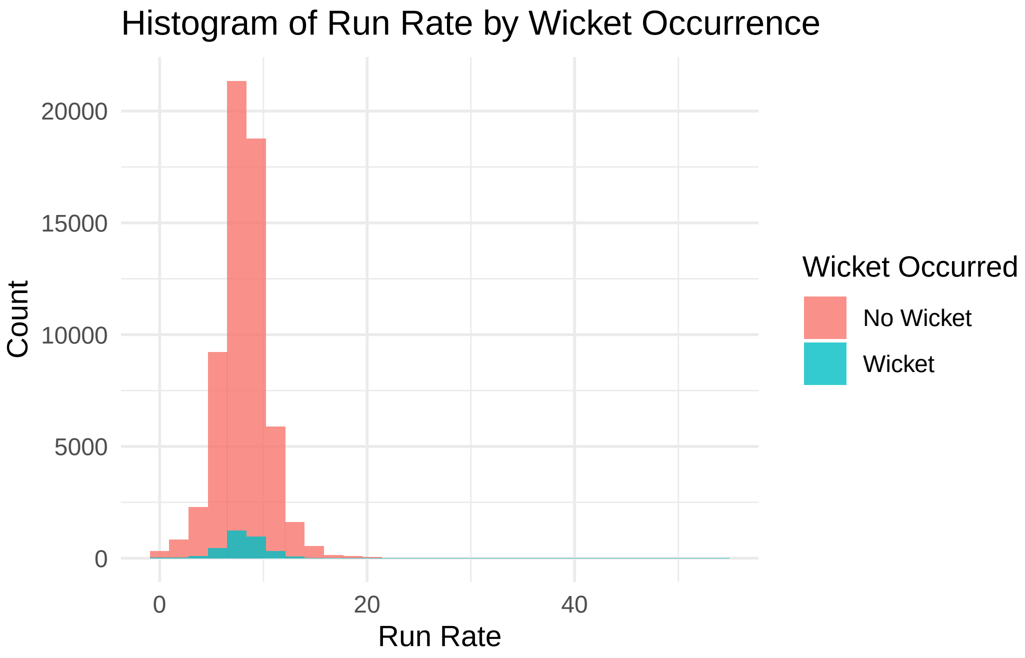
```
ggplot(bowling_batting_matchup_boundaries, aes(x = bowling_style, y = num_wickets, fill = ba
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(
       x = "Bowling Style",
       y = "Number of Wickets") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust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



```
ggplot(cleaned_data, aes(x = run_rate, fill = factor(wicket))) +
  geom_histogram(position = "identity", alpha = 0.8, bins = 30) +
  labs(
    x = "Run Rate",
    y = "Count",
    title = "Histogram of Run Rate by Wicket Occurrence"
  ) +
  scale_fill_discrete(name = "Wicket Occurred", labels = c("No Wicket", "Wicket")) +
  theme_minimal()
```


Histogram of Run Rate by Wicket Occurrence

3 Model

3.1 Model set-up

3.1.1 Model justification

4 Results

```
simple_glm_wicket_model <- readRDS(here("models/simple_glm_wicket_model.rds"))

#summary(simple_glm_wicket_model)
modelsummary(simple_glm_wicket_model)
```

```
complex_glm_wicket_model <- readRDS(here("models/complex_glm_wicket_model.rds"))
#summary(complex_glm_wicket_model)
modelsummary(complex_glm_wicket_model)
```

	(1)
(Intercept)	−3.569
	(0.048)
over	0.058
	(0.004)
Num.Obs.	51 587
AIC	20 589.7
BIC	20 607.4
Log.Lik.	−10 292.858
RMSE	0.22

	(1)
(Intercept)	−4.139
	(0.052)
over	0.005
	(0.004)
prev_over_wickets	1.852
	(0.031)
Num.Obs.	51 587
AIC	16 590.1
BIC	16 616.7
Log.Lik.	−8292.074
RMSE	0.21

```
overly_complex_glm_wicket_model <- readRDS(here("models/overly_complex_glm_wicket_model.rds"
#summary(overly_complex_glm_wicket_model)
modelsummary(overly_complex_glm_wicket_model)
```

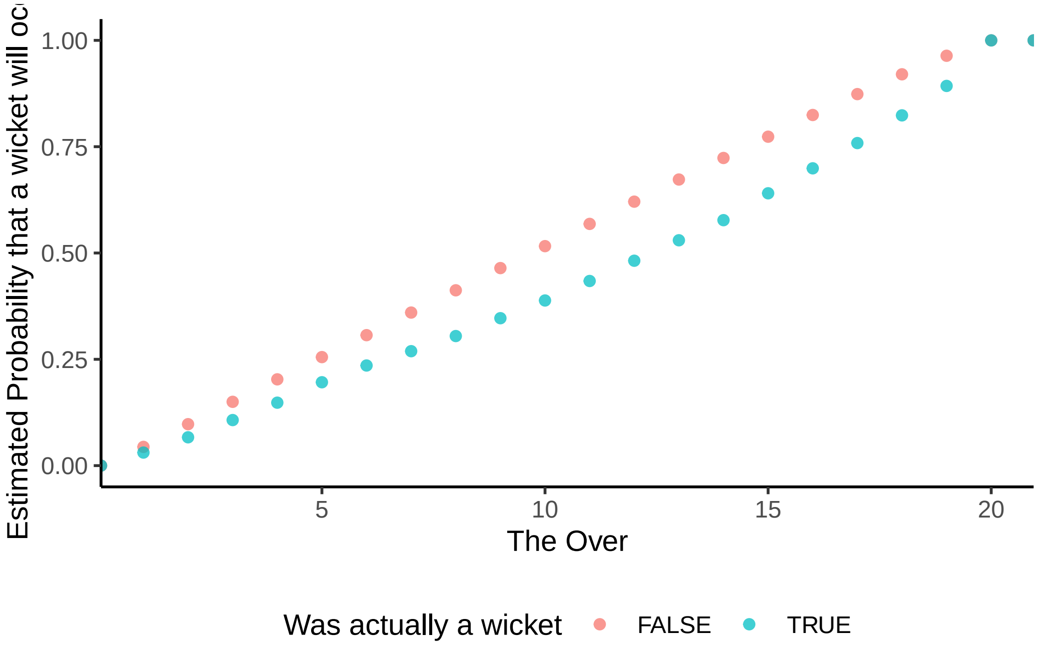
5 Simple Model Summary

```
overly_complex_glm_wicket_model <- complex_glm_wicket_model <- readRDS(here("models/overly_c

#summary(overly_complex_glm_wicket_model)
modelsummary(overly_complex_glm_wicket_model)
```

```
simple_glm_wicket_model_predictions <-
  predictions(simple_glm_wicket_model) |>
  as_tibble()

simple_glm_wicket_model_predictions |>
  mutate(wicket = factor(wicket)) |>
  ggplot(aes(x = over, y = estimate, color = wicket)) +
  stat_ecdf(geom = "point", alpha = 0.75) +
  labs(
    x = "The Over",
    y = "Estimated Probability that a wicket will occur",
    color = "Was actually a wicket"
  ) +
  theme_classic() +
  theme(legend.position = "bottom")
```

	(1)
(Intercept)	−4.316
	(0.192)
over	0.005
	(0.004)
prev_over_wickets	1.853
	(0.031)
batting_styleRight hand Bat	0.044
	(0.045)
bowling_styleLeft arm Fast medium	0.099
	(0.207)
bowling_styleLeft arm Medium	0.255
	(0.212)
bowling_styleLeft arm Medium fast	0.205
	(0.206)
bowling_styleLeft arm Wrist spin	0.158
	(0.235)
bowling_styleLegbreak	0.257
	(0.212)
bowling_styleLegbreak Googly	0.207
	(0.196)
bowling_styleRight arm Fast	0.190
	(0.192)
bowling_styleRight arm Fast medium	0.211
	(0.197)
bowling_styleRight arm Medium	0.109
	(0.193)
bowling_styleRight arm Medium fast	0.165
	(0.204)
bowling_styleRight arm Offbreak	0.104
	(0.200)
bowling_styleSlow Left arm Orthodox	0.011
	(0.202)
Num.Obs.	51 587
AIC	16 605.8
BIC	16 747.4
Log.Lik.	−8286.914
RMSE	0.21

	(1)
(Intercept)	−4.316
	(0.192)
over	0.005
	(0.004)
prev_over_wickets	1.853
	(0.031)
batting_styleRight hand Bat	0.044
	(0.045)
bowling_styleLeft arm Fast medium	0.099
	(0.207)
bowling_styleLeft arm Medium	0.255
	(0.212)
bowling_styleLeft arm Medium fast	0.205
	(0.206)
bowling_styleLeft arm Wrist spin	0.158
	(0.235)
bowling_styleLegbreak	0.257
	(0.212)
bowling_styleLegbreak Googly	0.207
	(0.196)
bowling_styleRight arm Fast	0.190
	(0.192)
bowling_styleRight arm Fast medium	0.211
	(0.197)
bowling_styleRight arm Medium	0.109
	(0.193)
bowling_styleRight arm Medium fast	0.165
	(0.204)
bowling_styleRight arm Offbreak	0.104
	(0.200)
bowling_styleSlow Left arm Orthodox	0.011
	(0.202)
Num.Obs.	51 587
AIC	16 605.8
BIC	16 747.4
Log.Lik.	−8286.914
RMSE	0.21

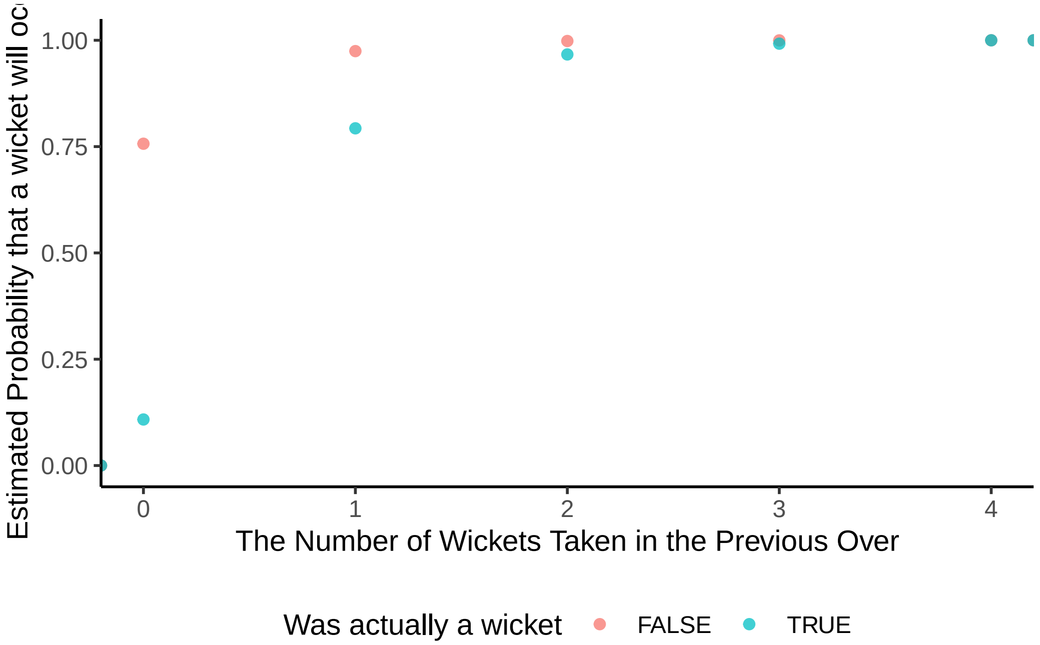


```
test_data_simple <- test_data
predictions <- predict(simple_glm_wicket_model, newdata = test_data_simple, type = "response

test_data_simple$predicted_wicket_prob <- predictions
test_data_simple <- test_data_simple %>%
  mutate(predicted_wicket = predicted_wicket_prob >= 0.5) %>%
  mutate(correct_prediction = predicted_wicket == wicket)

summary_results <- test_data_simple %>% group_by(wicket) %>%
  summarise(
  correct = sum(correct_prediction),
  incorrect = sum(!correct_prediction)
)

summary_results
```

```
# A tibble: 2 x 3
  wicket correct incorrect
  <lgl>    <int>     <int>
1 FALSE    12253         0
2 TRUE         0       644
```

5.1 Complex Model Summary

```
complex_glm_wicket_model_predictions <-
  predictions(complex_glm_wicket_model) |>
  as_tibble()

complex_glm_wicket_model_predictions |>
  mutate(wicket = factor(wicket)) |>
  ggplot(aes(x = prev_over_wickets, y = estimate, color = wicket)) +
  stat_ecdf(geom = "point", alpha = 0.75) +
  labs(
    x = "The Number of Wickets Taken in the Previous Over",
    y = "Estimated Probability that a wicket will occur",
    color = "Was actually a wicket"
  ) +
  theme_classic() +
  theme(legend.position = "bottom")
```



```
test_data_complex <- test_data
predictions <- predict(complex_glm_wicket_model, newdata = test_data_complex, type = "respons

test_data_simple$predicted_wicket_prob <- predictions
test_data_simple <- test_data_simple %>%
  mutate(predicted_wicket = predicted_wicket_prob >= 0.5) %>%
  mutate(correct_prediction = predicted_wicket == wicket)

summary_results <- test_data_simple %>% group_by(wicket) %>%
  summarise(
  correct = sum(correct_prediction),
  incorrect = sum(!correct_prediction)
)

summary_results
```

```
# A tibble: 2 x 3
  wicket correct incorrect
  <lgl>    <int>     <int>
1 FALSE    12229        24
2 TRUE        17       627
```

6 Discussion

6.1 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.

6.2 Second discussion point

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

6.3 Third discussion point

6.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

```
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.
```

```
bowling_batting_role_matchup_boundaries
```

```
# A tibble: 117 x 4
# Groups:   bowling_style [13]
   bowling_style          batter_playing_role num_wickets num_balls
   <chr>                  <chr>                     <int>     <int>
 1 Left arm Fast          Allrounder                   10       115
 2 Left arm Fast          Batter                        2        95
 3 Left arm Fast          Batting Allrounder            2        79
 4 Left arm Fast          Bowler                        4        37
 5 Left arm Fast          Bowling Allrounder            2        18
 6 Left arm Fast          Middle order Batter           4        89
 7 Left arm Fast          Opening Batter                4       174
 8 Left arm Fast          Top order Batter              9       137
 9 Left arm Fast          Wicketkeeper Batter          12       190
10 Left arm Fast medium Allrounder                     25       503
# i 107 more rows
```

B Model details

B.1 Posterior predictive check

C References

What to cite: - cricketdata - ESPNCricinfo - Cricsheet - All tidyverse packages used