MATHS 7107 Data Taming
Assignment 02

This assignment is created by Possakorn Kittipipatthanapong (Student id: a1873765). And, it consists of 5 sections following the question below:

Question One: Reading and Cleaning

```
library(tidyverse)
library(tinytex)
library(knitr)
library(janitor)
library(readr)
library(float)
library(kableExtra)

df <- read_csv("C:/Users/possa/OneDrive - University of Adelaide/script_courses/c_2023_01_comp sci_7107/

kable(df[1:5,], format = "latex", caption = "Ashes Data")
```

(a) For our analysis, the subjects are not the cricketers themselves, but each batting innings they participated in. In order to make the data tidy:

```
# set the pattern
pattern_detail = "Batting at number (.*?) scored (.*?) runs from (.*?) balls including (.*?) fours and
pattern_innings = "Test.(.*?)_Innings_(.?)"

df_prep <- df %>%
  gather(key = "innings_test", value = "detail", c(-batter, -team, -role)) %>%
  mutate(str_match(innings_test, pattern_innings) %>%
           as_tibble(.name_repair = ~ c("matched_2", "test_match", "innings"))
         ) %>%
```

Table 1: Ashes Data

batter	team	role	Test 1_Innings_1	T
Ali	Eng	allrounder	Batting at number 8 scored 0 runs from 5 balls including 0 fours and 0 sixes.	B
Anderson	England	bowl	Batting at number 11 scored 3 runs from 19 balls including 0 fours and 0 sixes.	B
Archer	England	bowl	Batting at number NA scored NA including NA fours and NA sixes.	B
Bairstow	England	wicketkeeper	Batting at number 7 scored 8 runs from 35 balls including 1 fours and 0 sixes.	B
Bancroft	Aus	bat	Batting at number 1 scored 8 runs from 25 balls including 2 fours and 0 sixes.	B

Table 2: Shown the data tidy with 5 rows

batter	team	role	test_match	innings	number	scores	balls	fours	sixes
Ali	Eng	allrounder	1	1	8	0	5	0	0
Anderson	England	bowl	1	1	11	3	19	0	0
Archer	England	bowl	1	1	NA	NA	NA	NA	NA
Bairstow	England	wicketkeeper	1	1	7	8	35	1	0
Bancroft	Aus	bat	1	1	1	8	25	2	0

```
  mutate(str_match(detail, pattern_detail) %>%
           as_tibble(.name_repair = ~ c("matched", "number", "scores", "balls", "fours", "sixes"))
        ) %>%
  select(-matched, -matched_2, -innings_test, -detail)

kable(df_prep[1:5,], format = "latex", caption = "Shown the data tidy with 5 rows")
```

(b) Recode the data to make it 'tame'

```
# check type of dataframe
library(inspectdf)

inspect_cat(df_prep)
```

```
## # A tibble: 10 x 5
##    col_name     cnt common    common_pcnt levels
##    <chr>      <int> <chr>           <dbl> <named list>
##  1 balls        101 <NA>            33.2  <tibble [101 x 3]>
##  2 batter        31 Ali              3.23 <tibble [31 x 3]>
##  3 fours         17 <NA>            33.2  <tibble [17 x 3]>
##  4 innings        2 1               50    <tibble [2 x 3]>
##  5 number        12 <NA>            33.2  <tibble [12 x 3]>
##  6 role          10 bat             25.8  <tibble [10 x 3]>
##  7 scores        71 <NA>            33.2  <tibble [71 x 3]>
##  8 sixes          7 0               61.0  <tibble [7 x 3]>
##  9 team           4 Australia       48.4  <tibble [4 x 3]>
## 10 test_match     5 1               20    <tibble [5 x 3]>
```

```
show_plot(inspect_cat(df_prep))
```

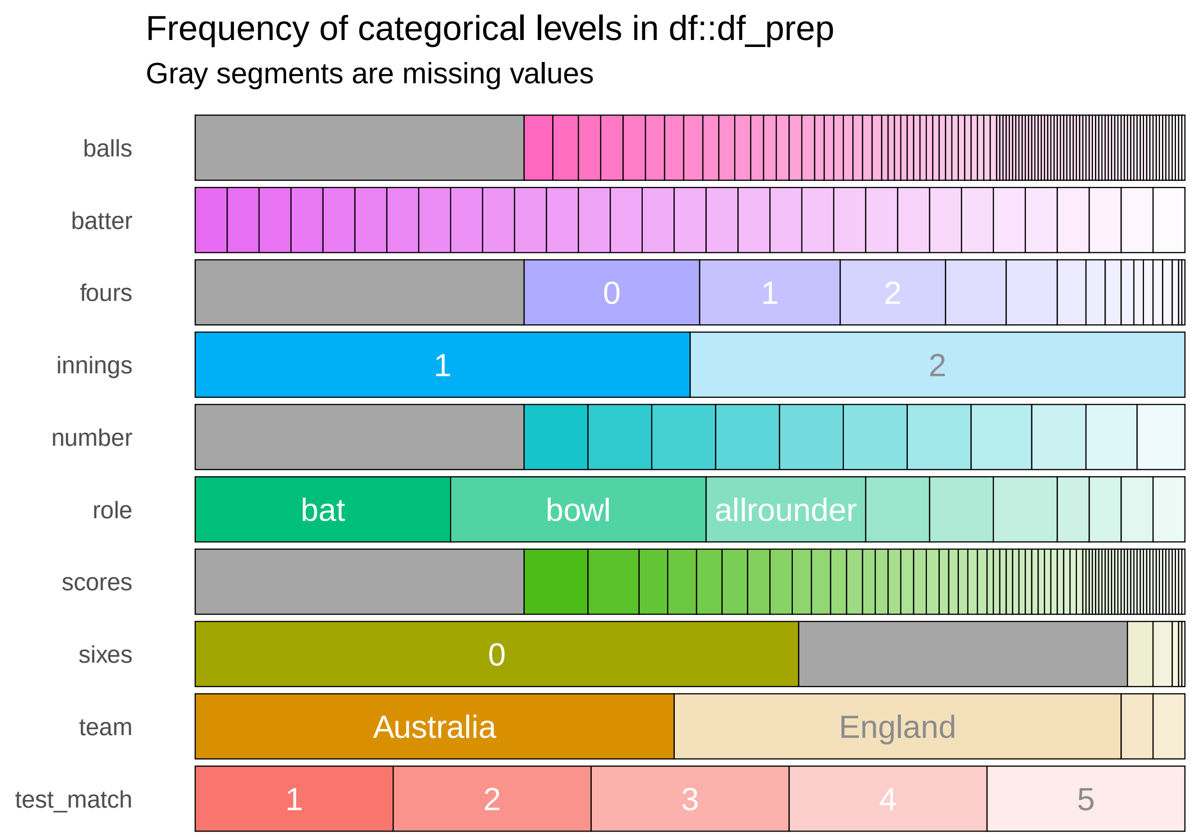

Figure 1: The inspection of category variable

```
# Adjust column type
df_prep_2 <- df_prep %>%
  mutate(
    # Ensure all categorical variables with a small number of levels are coded as factors
    team = as.factor(team),
    role = as.factor(role),
    innings = as.factor(innings),
    # Ensure all categorical variables with a large number of levels are coded as characters - no need
    # Ensure all quantitative variables are coded as integers
    number = as.integer(number),
    scores = as.integer(scores),
    balls = as.integer(balls),
    fours = as.integer(fours),
    sixes = as.integer(sixes)
    # Ensure all quantitative variables are coded as numeric - no need b/c all integer
         )

str(df_prep_2)
```

```
## tibble [310 x 10] (S3: tbl_df/tbl/data.frame)
##  $ batter    : chr [1:310] "Ali" "Anderson" "Archer" "Bairstow" ...
##  $ team      : Factor w/ 4 levels "Aus","Australia",..: 3 4 4 4 1 4 4 4 2 4 ...
##  $ role      : Factor w/ 10 levels "all-rounder",..: 3 7 7 10 4 8 4 4 8 7 ...
```

```
##  $ test_match: chr [1:310] "1" "1" "1" "1" ...
##  $ innings   : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
##  $ number    : int [1:310] 8 11 NA 7 1 10 1 5 9 NA ...
##  $ scores    : int [1:310] 0 3 NA 8 8 29 133 5 5 NA ...
##  $ balls     : int [1:310] 5 19 NA 35 25 67 312 10 10 NA ...
##  $ fours     : int [1:310] 0 0 NA 1 2 2 17 1 1 NA ...
##  $ sixes     : int [1:310] 0 0 NA 0 0 0 0 0 0 NA ...
```

(c) Clean the data; recode the factors using fct_recode() such that there are no typographical errors in the team names and player roles

Reference the role following the question guideline

```
# typographical errors - check

count(df_prep_2, team)
```

```
## # A tibble: 4 x 2
##   team          n
##   <fct>     <int>
## 1 Aus          10
## 2 Australia   150
## 3 Eng          10
## 4 England     140
```

```
count(df_prep_2, role)
```

```
## # A tibble: 10 x 2
##    role             n
##    <fct>        <int>
##  1 all-rounder     10
##  2 all rounder     10
##  3 allrounder      50
##  4 bat             80
##  5 batsman         20
##  6 batting         10
##  7 bowl            80
##  8 bowler          20
##  9 bowling         10
## 10 wicketkeeper    20
```

```
# data manipulation
df_prep_final <- df_prep_2 %>%
  mutate(team =
           team %>%
           fct_recode(
             Australia = "Aus",
             England = "Eng"
           ),
         role =
           role %>%
           fct_recode(
```

Table 3: Shown the data tidy after recode the with 5 rows

batter	team	role	test_match	innings	number	scores	balls	fours	sixes
Ali	England	all-rounder	1	1	8	0	5	0	0
Anderson	England	bowler	1	1	11	3	19	0	0
Archer	England	bowler	1	1	NA	NA	NA	NA	NA
Bairstow	England	wicketkeeper	1	1	7	8	35	1	0
Bancroft	Australia	batter	1	1	1	8	25	2	0

```
             "all-rounder" = "allrounder",
             "all-rounder" = "all rounder",
             batter = "bat",
             batter = "batting",
             batter = "batsman",
             bowler = "bowl",
             bowler = "bowling"
           )
         )
# typographical errors - recheck
kable(df_prep_final[1:5,], format = "latex", caption = "Shown the data tidy after recode the with 5 rows
```

```
count(df_prep_final, team)
```

```
## # A tibble: 2 x 2
##   team          n
##   <fct>     <int>
## 1 Australia   160
## 2 England     150
```

```
count(df_prep_final, role)
```

```
## # A tibble: 4 x 2
##   role             n
##   <fct>        <int>
## 1 all-rounder     70
## 2 batter         110
## 3 bowler         110
## 4 wicketkeeper    20
```

Question Two: Univariate Analysis

(a) Produce a histogram of all scores during the series.

```
df_prep_final %>%
  gather(key = "output", value = "value", c(-batter, -team, -role, -innings,-test_match, -number)) %>%
  ggplot(aes(x = value, fill = output)) +
  geom_histogram(alpha=.6, width=.6) +
  facet_wrap(vars(output), scales = "free") +
```

```
  labs( x = 'value',
        y = 'n()'
        )
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

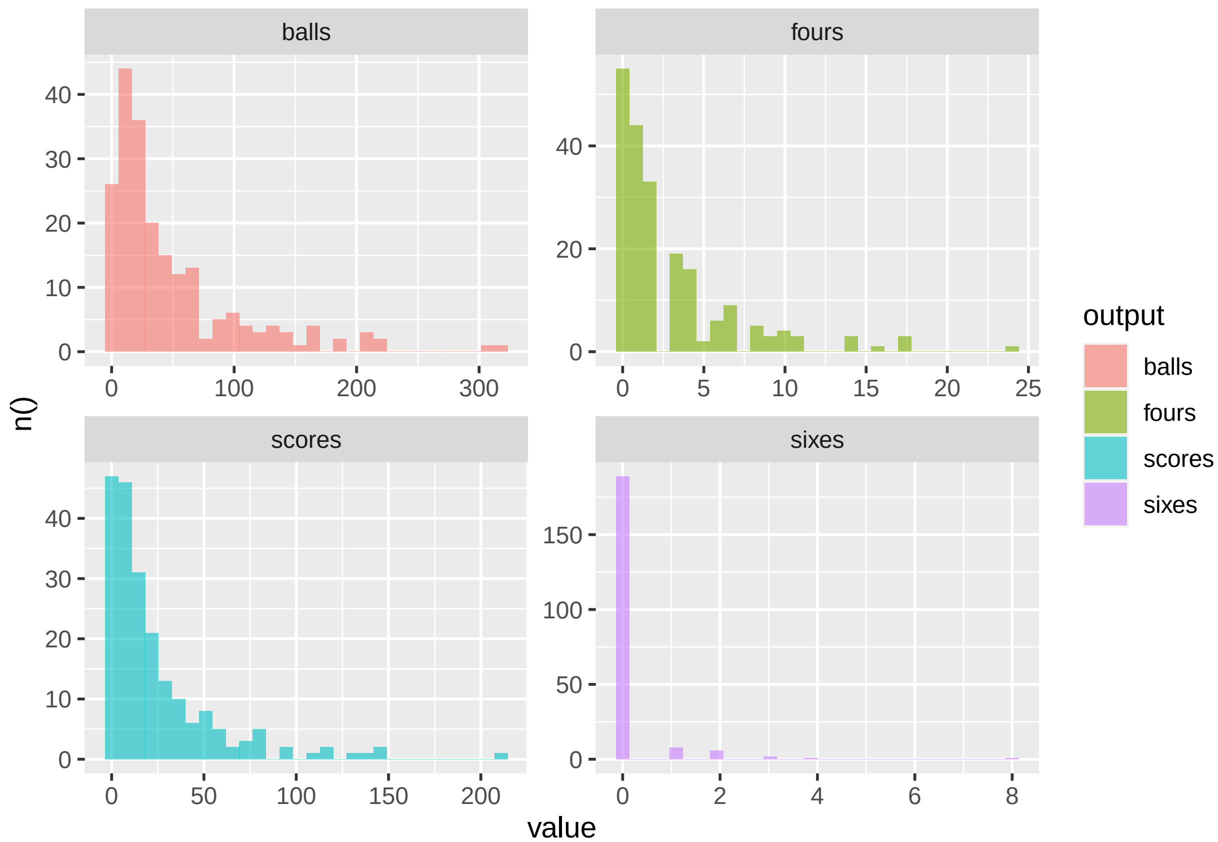

Figure 2: Histrograms for each value in the series including the scores

(b) Describe the distribution of scores, considering shape, location spread and outliers.

```
# show the box_plot to define the distributions
df_prep_final %>%
  gather(key = "output", value = "value", c(-batter, -team, -role, -innings,-test_match, -number)) %>%
  ggplot(aes(x = value, fill = output)) +
  geom_boxplot(alpha=.6, width=.6) +
  facet_wrap(vars(output), scales = "free") +
  labs( x = 'value',
        y = ''
        )
```

Table 4: Checking 5 numbers and related values

col_name	min	q1	median	mean	q3	max	sd	pcnt_na	hist
scores	0	4.0	12	23.9420290	30.5	211	31.6986190	33.22581	[0, 10) , [10, 20) , [20, 30) , [30, 4
balls	1	12.5	26	47.9516908	61.5	319	55.6294745	33.22581	[0, 20) , [20, 40) , [40, 60) , [60, 8
fours	0	0.0	2	2.9613527	4.0	24	3.8840543	33.22581	[0, 1) , [1, 2) , [2, 3) , [3, 4) , [4, 5
sixes	0	0.0	0	0.1835749	0.0	8	0.7727902	33.22581	[0, 0.5) , [0.5, 1) , [1, 1.5) , [1.5, 2

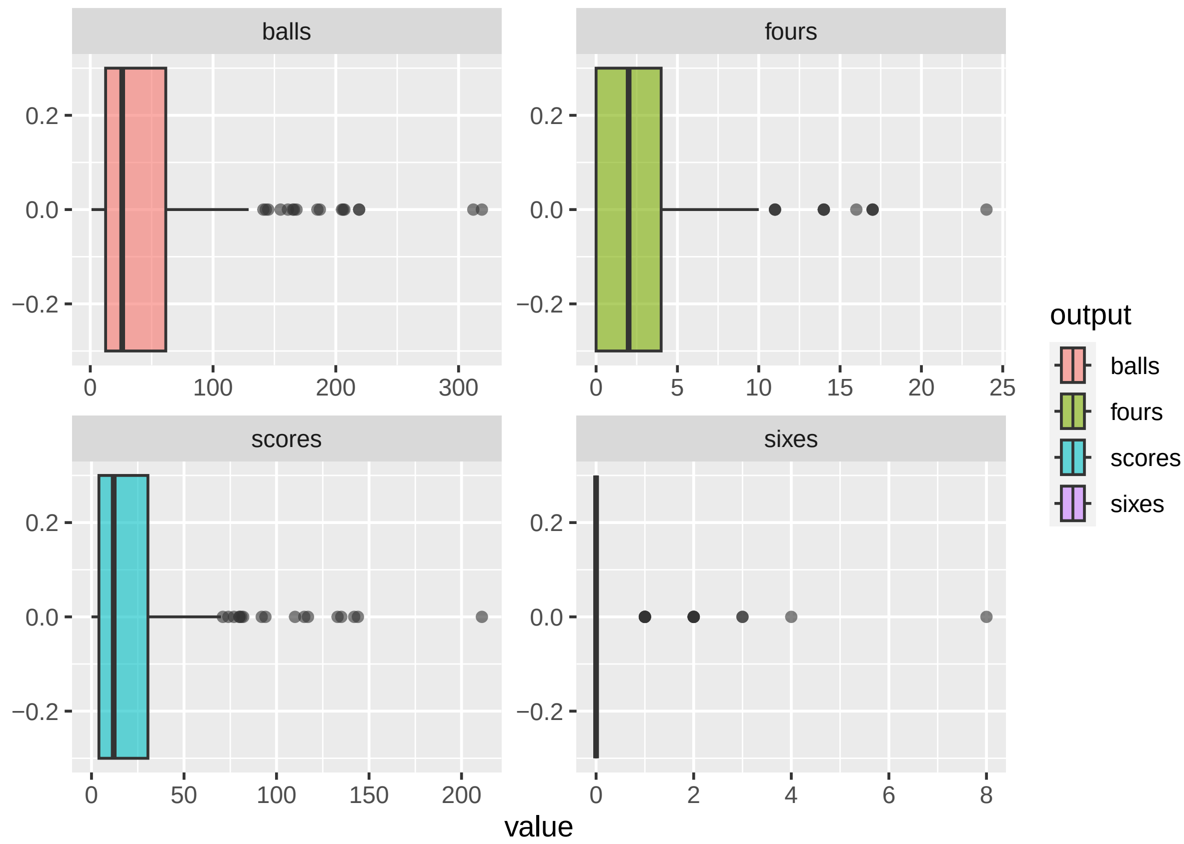

Figure 3: Boxplot for each value in the series including the scores

```
check_5number <- df_prep_final %>%
  select(-test_match,-number) %>%
  inspect_num()

kable(check_5number, format = "latex", caption = "Checking 5 numbers and related values")
```

Answer: In term of shape, the scores values represent the right-skewness. Related to the locaion, mean and median equal to 23.942 and 12.000. For Spread, this series has the SD and IQR equal to 31.698 and 26.500. For Outlier, there are some outliers above 70.25.

(c) Produce a bar chart of the teams participating in the series, with different colours for each team. Noting that each player is represented by 10 rows in the data frame, how many players were used by each team in the series?

```
df_prep_final %>%
  group_by(team) %>%
  summarise(
    active_player = n_distinct(batter[!is.na(number)] )
  ) %>%
  ggplot(aes(x = team, y = active_player, fill = team)) +
  geom_bar(stat="identity",  alpha=.6, width=.6) +
  coord_flip() +
  labs( y = 'Active Player', x = 'Team') +
  guides(fill = "none")
```

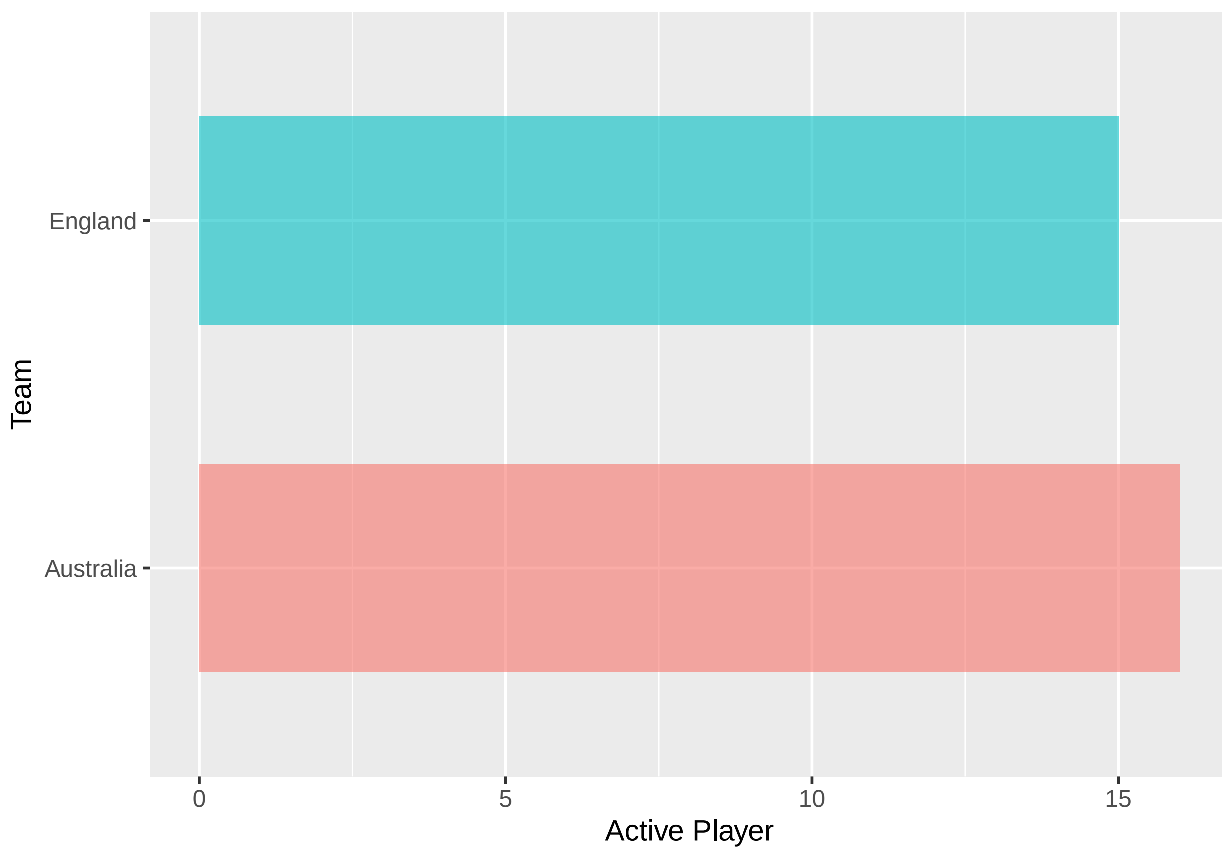

Figure 4: Bar Chart shown the used player for each team in the series

Question Three: Scores for each team

(a) Using ggplot, produce histograms of scores during the series, faceted by team.

```
df_prep_final %>%
  ggplot(aes(x = scores, fill = team)) +
  geom_histogram(alpha=.6, width=.6) +
  facet_grid(rows = vars(team)) +
  labs( x = 'Score',
        y = ''
        ) +
  guides(fill = "none")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

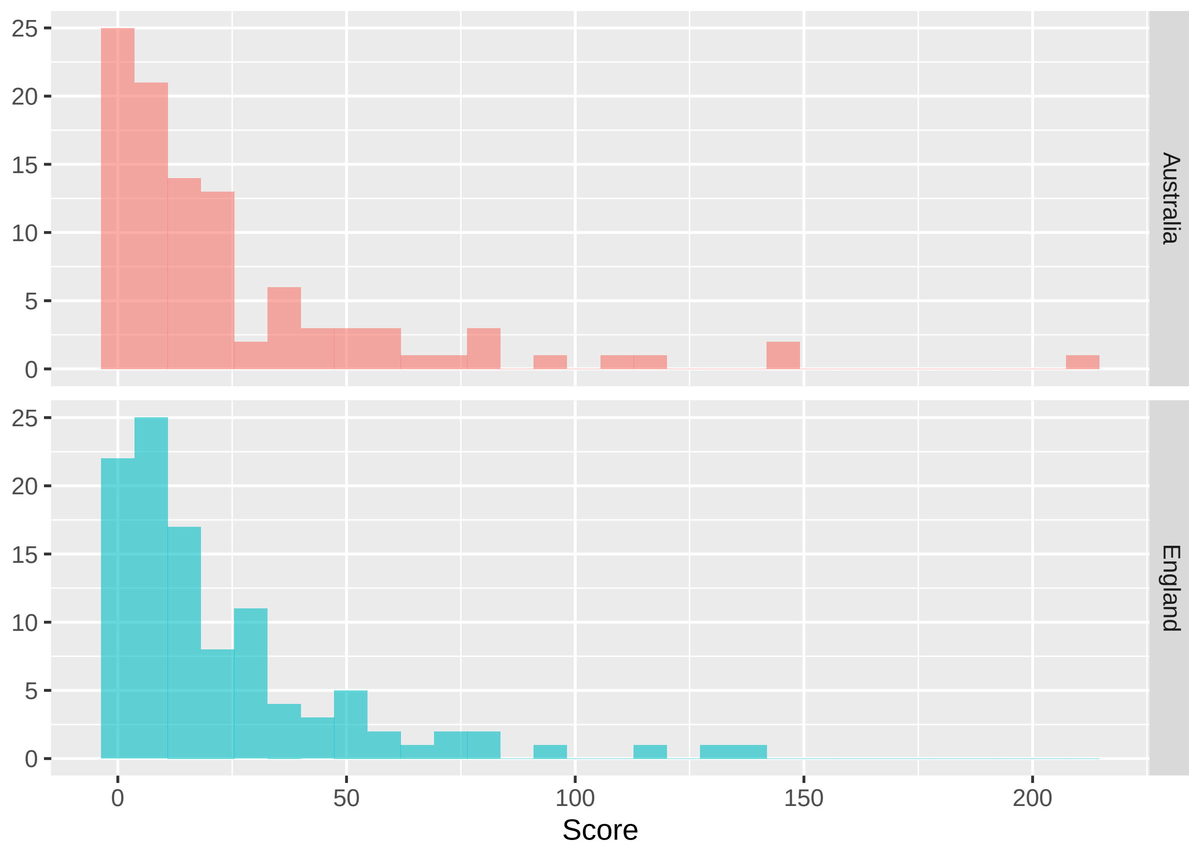

Figure 5: Histograms of scores during the series for each team

(b) Produce side-by-side boxplots of scores by each team during the series.

```
df_prep_final %>%
  ggplot(aes(x = scores, fill = team)) +
  geom_boxplot(alpha=.6, width=.6) +
  facet_grid(rows = vars(team)) +
  labs( x = 'Score',
        y = ''
```

Table 5: Checking 5 numbers for each teams

team	col_name	min	q1	median	mean	q3	max	sd	pcnt_na	hist
England	scores	0	4	12	22.55660	30	135	27.50587	29.33333	[0, 10) , [10, 20) , [20, 30) , [3
Australia	scores	0	4	12	25.39604	33	211	35.65560	36.87500	[0, 10) , [10, 20) , [20, 30) , [3

```
        ) +
  guides(fill = "none")
```

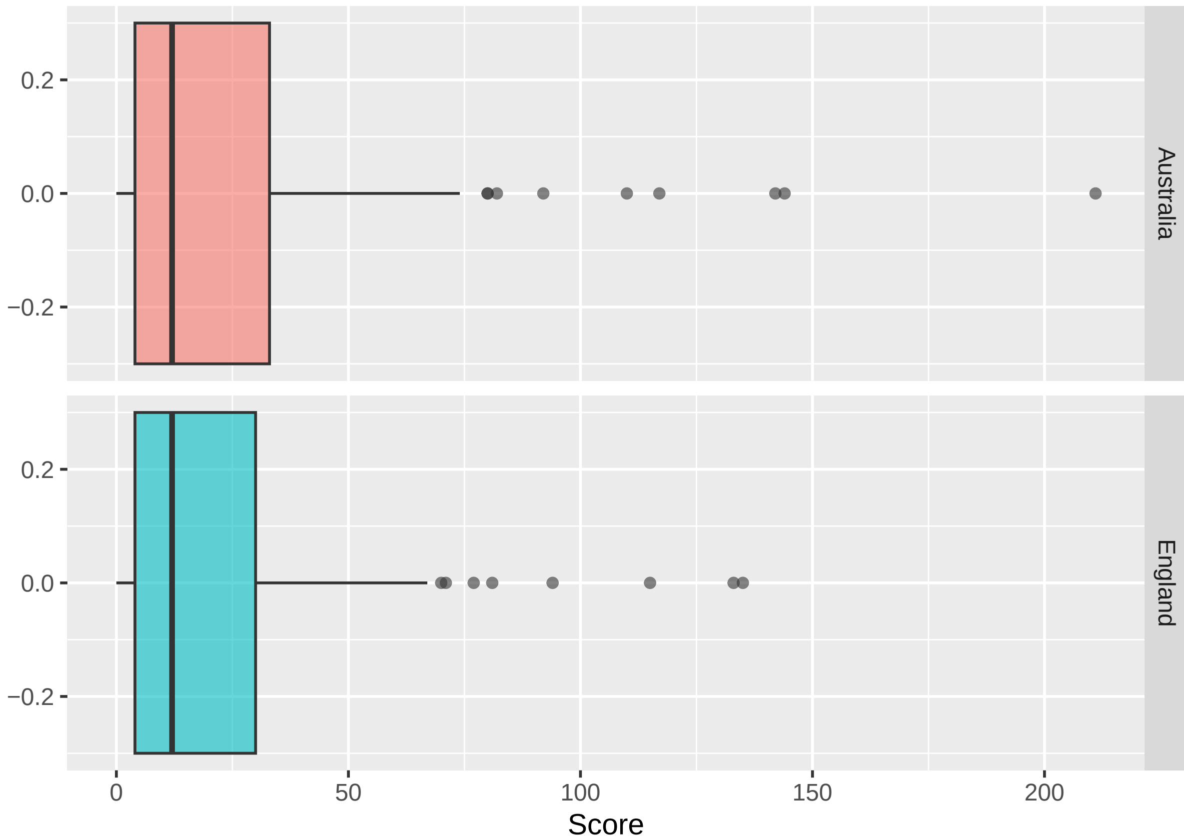

Figure 6: boxplots of scores by each team during the series

```
check_5number_2 <- df_prep_final %>%
  group_by(team) %>%
  inspect_num() %>%
  filter(col_name == "scores")
kable(check_5number_2, format = "latex", caption = "Checking 5 numbers for each teams")
```

(c) Compare the distributions of scores by each team during the series, considering shape, location, spread and outliers, and referencing the relevant plots. Which team looks to have had a higher variablility of scores?

Answer: In term of shape, both of team represent the unimodel and right-skewed distribution. For location, median score of the players in each team present the equivalent trend. On the other hand, mean of Australia slightly higher than England (25.396 > 22.556). For spread, As Australia also show the higher distribution related to the SD compared to England (33 > 30). For outlier, Australia have a few outlier above 68.896 instead of 61.556 for England.

Question Four: Scoring rates

(a) Produce a scatterplot of scores against number of balls.

```
df_prep_final %>%
  ggplot(aes(x = balls, y = scores)) +
  geom_point()+
  geom_smooth(col = 'red')
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

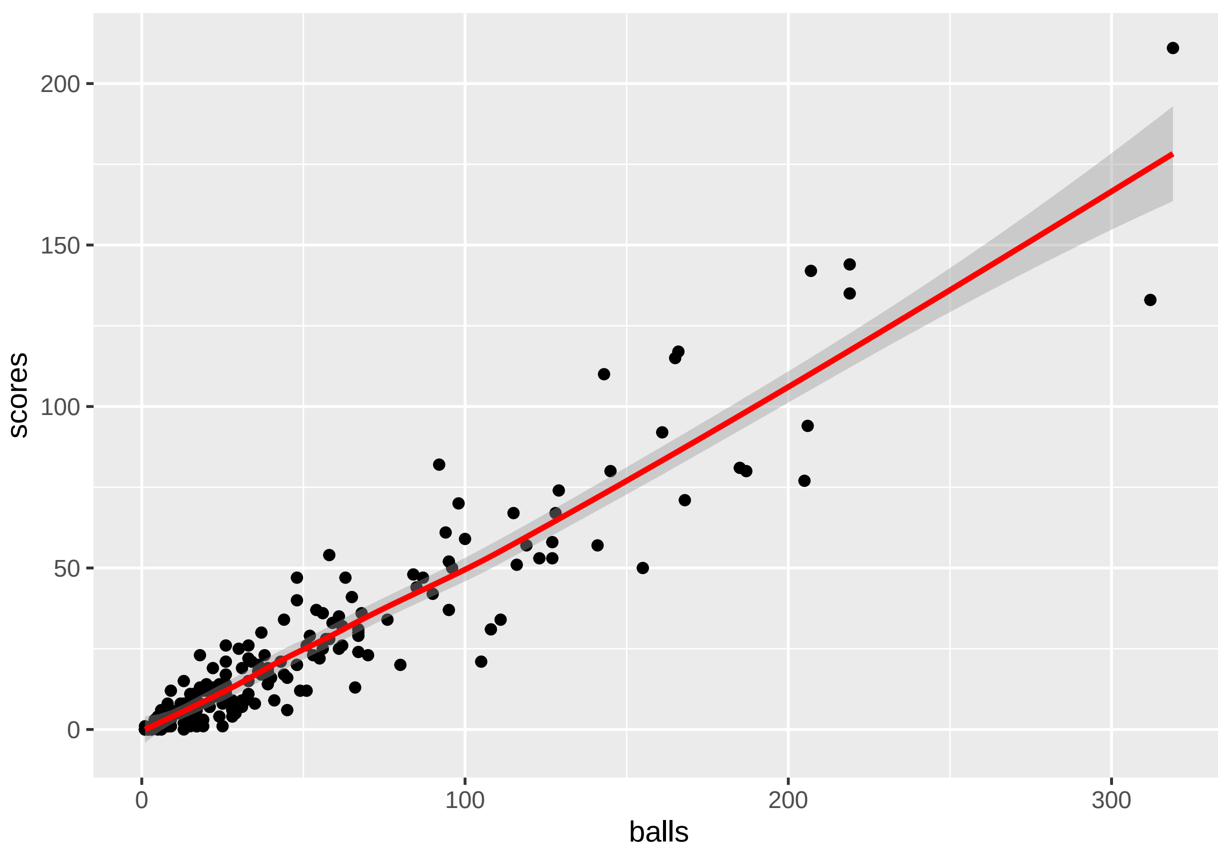

Figure 7: The scatterplot between number of balls and scores

(b) Describe the relationship between score and number of balls. Are players who face more balls likely to score more runs?

Answer: the relationship between score and number of balls represent in moderate linear relationship. For Player who face more balls, they likely to score more runs.

(c) Compute a new variable, scoring_rate, defined as the number

```
df_prep_final <-
  df_prep_final %>%
  mutate(scoring_rate = scores / balls)
```

(d) Is there a relationship between scoring rate and number of balls? Are players who face more balls likely to score runs more quickly?

```
df_prep_final %>%
  ggplot(aes(x = balls, y = scoring_rate)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

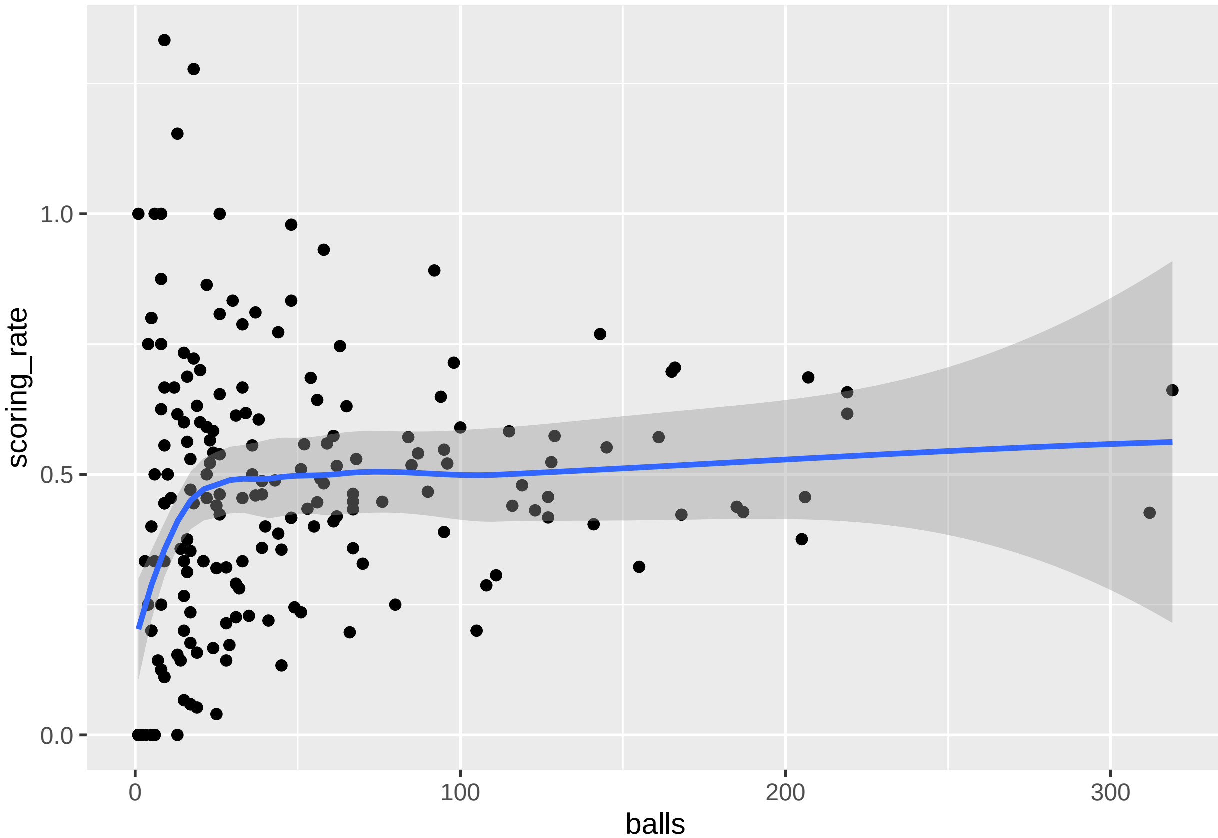

Figure 8: The scatterplot shown relationship between scoring rate and number of balls

Answer: the relationship between score and number of balls don't represent the linear relationship. After the certain amount of balls, the scoring approach the stable value. Therefore, Player who face more balls. It doesn't mean they will get the higher scoring rate.

Question Five: Teams' roles

(a) Produce a bar chart of the number of players on each team participating in the series, with segments coloured by the players' roles.

```
df_prep_final %>%
  group_by(team, role) %>%
  summarise(
    active_player = n_distinct(batter[!is.na(number)] )
  ) %>%
  ggplot(aes(x = team, y = active_player, fill = role)) +
  geom_bar(stat="identity",  alpha=.6, width=.6) +
  coord_flip() +
  labs(y = "Number of unique player",
       x = '')
```

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

Table 6: a contingency table of the proportion of players from each team

team	all-rounder	batter	bowler	wicketkeeper	Total
Australia	13%	44%	38%	6%	100%
England	33%	27%	33%	7%	100%

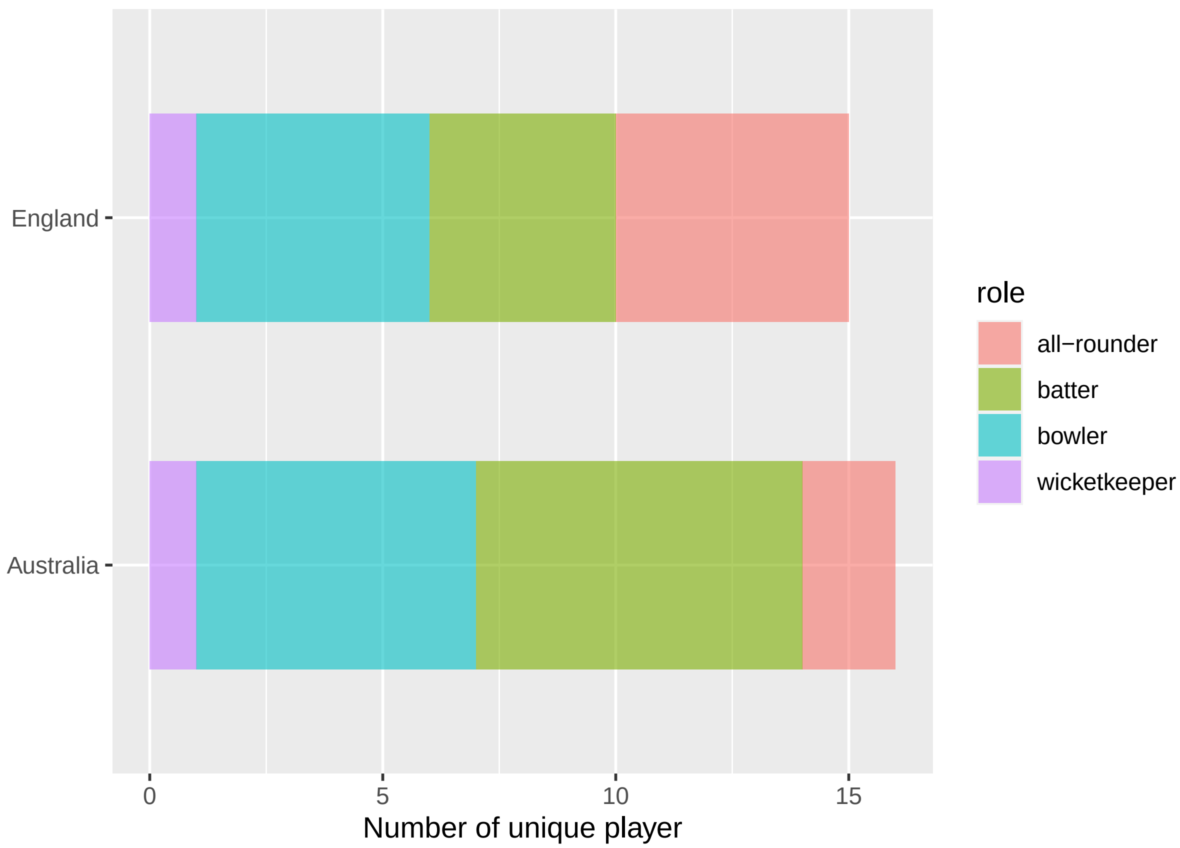

Figure 9: The bar chart of the number of players on each team participating in the series, with segments

(b) Produce a contingency table of the proportion of players from each team who play in each particular role.

```
# Create contingency table
contingency <- df_prep_final %>%
  tabyl(team, role) %>%
  adorn_percentages("row") %>%
  adorn_totals(c("col")) %>%
  adorn_pct_formatting(rounding = "half up", digits = 0) %>%
  tibble()

kable(contingency, format = "latex", caption = "a contingency table of the proportion of players from ea
```

```r
# Create stack 100 chart
df_prep_final %>%
  group_by(team, role) %>%
  summarise(
    active_player = n_distinct(batter[!is.na(number)] )
  ) %>%
  ggplot(aes(x = team, y = active_player, fill = role)) +
  geom_bar(stat="identity",position="fill",  alpha=.6, width=.6) +
  coord_flip() +
  labs(y = "Proportion from total",
       x = '')
```

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

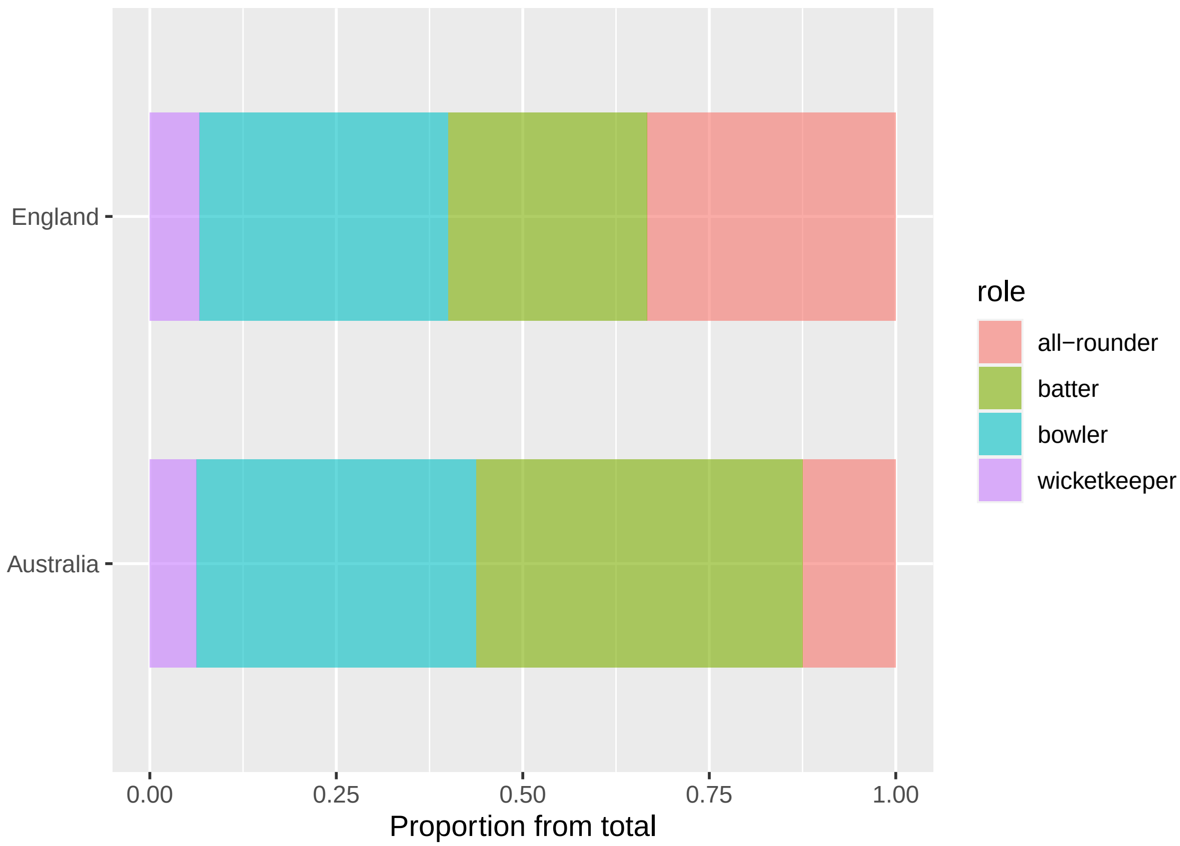

Figure 10: The stack 100 bar chart of the number of players on each team participating in the series, with segments

(c) Using these two figures, state which team is made up of a larger proportion of batters, and which team contains a larger proportion of all-rounders.

Answer: For Australia, there are larger proportion of batters and less proportion of all-rounders compared to England.