Analysis of the 2017–2018 Ashes Men's Cricket Series (Australia vs England)

Introduction:

This report analyses batting performances during the 2017–18 Men's Ashes cricket series between Australia and England. Using R, we explore and clean the dataset, transform it into a tidy format, and conduct both univariate and bivariate analyses. The goal is to extract insights about the teams' scoring trends, player performances, and batting behaviour throughout the series. The data includes each player's batting results across up to ten innings (two per Test match). Each record describes the player's batting position, runs scored, and number of balls faced, all embedded in text sentences.

Reading and Cleaning:

This section focuses on loading, tidying, and cleaning the *Ashes 2017–18* cricket dataset to prepare it for analysis by organising the data, extracting key variables, and ensuring all values are correctly coded.

loading the Ashes dataset into RStudio & convert it into long format:

R Code Input:

```
#load the data from Excel
library(readxl)

#location of file
setwd("C:/Users/PC/Downloads")

#import the Excel file
ashes <- read.csv("ashes.csv")
View(ashes)
```

```
#convert to long format
install.packages("tidyr")
library(tidyr)
ashes_long <- ashes %>%
  pivot_longer(
    cols = starts_with("Test"),
    names_to = "innings",
    values_to = "performance"
  )
View(ashes_long)
head(ashes_long)
```

R Output:

```
# A tibble: 6 × 5
 batter team    role          innings               performance
  <chr>  <chr>   <chr>         <chr>                 <chr>
1 Ali    England allrounder Test.1..Innings.1 Batting at number 6, scored 38 runs from 102 balls including 2 fours and 1…
2 Ali    England allrounder Test.1..Innings.2 Batting at number 6, scored 40 runs from 64 balls including 6 fours and 0 …
3 Ali    England allrounder Test.2..Innings.1 Batting at number 6, scored 25 runs from 57 balls including 2 fours and 0 …
4 Ali    England allrounder Test.2..Innings.2 Batting at number 7, scored 2 runs from 20 balls including 0 fours and 0 s…
5 Ali    England allrounder Test.3..Innings.1 Batting at number 7, scored 0 runs from 2 balls including 0 fours and 0 si…
6 Ali    England allrounder Test.3..Innings.2 Batting at number 7, scored 11 runs from 56 balls including 2 fours and 0 …
```

- 270 rows and 6 columns (batter, team, role, Test, Innings, performance)
- Each row now represents one player in one innings, instead of one player with many columns.

Explanation:

We changed the dataset from wide (many columns per player) to long (one row per batting innings).

This is essential because each observation (one batting performance) should have its own row, following tidy data principles.

Creating 3 new columns:

We will extract the batting number, score, and balls faced from the performance text using str_match() to create three separate numeric columns.

R Code Input:

```
library(stringr)

library(dplyr)

ashes_long <- ashes_long %>%

  mutate(

    parts = str_match(performance, "Batting at number ([0-9]+).*scored ([0-
9]+).*from ([0-9]+)"),

    batting_number = parts[,2],

    score = parts[,3],

    balls_faced = parts[,4]

  ) %>%

  select(batter, team, role, innings, performance, batting_number, score,
balls_faced)

head(ashes_long)
```

R Output:

```
# A tibble: 6 × 8

batter  team   role       innings   performance                                    batting_number    score    balls_faced

 <chr>  <chr>  <chr>      <chr>        <chr>                                          <chr>           <chr>      <chr>

1 Ali   England allrounder Test.1..Innings.1 Batting at number 6, scored 38 runs from …   6              38         102

2 Ali   England allrounder Test.1..Innings.2 Batting at number 6, scored 40 runs from …   6              40          64

3 Ali   England allrounder Test.2..Innings.1 Batting at number 6, scored 25 runs from …   6              25          57

4 Ali   England allrounder Test.2..Innings.2 Batting at number 7, scored 2 runs from 2…    7              2           20

5 Ali   England allrounder Test.3..Innings.1 Batting at number 7, scored 0 runs from 2…    7              0            2

6 Ali   England allrounder Test.3..Innings.2 Batting at number 7, scored 11 runs from …    7             11           56
```

Explanation:

We used the str_match() function to extract numeric information from each text sentence and create three new numeric columns:

- batting number: Player’s batting position

- score: Runs scored
- balls: Balls faced

This converts the text into usable numeric variables for statistical analysis.

Recoding variables:

We will assign the correct data types to each variable.

R Code Input:

```
ashes_long <- ashes_long %>%
  mutate(
    team = as.factor(team),
    role = as.factor(role),
    innings = as.factor(innings),
    batter = as.character(batter),
    performance = as.character(performance),
    batting_number = as.integer(batting_number),
    score= as.integer(score),
    balls_faced = as.integer(balls_faced)
    )
head(ashes_long)
```

R Output:

```
 # A tibble: 6 × 8
 batter  team    role         innings             performance                                         batting_number   score   balls_faced
 <chr>   <fct>   <fct>        <fct>               <chr>                                               <int>            <int>   <int>
 1 Ali   England allrounder Test.1..Innings.1 Batting at number 6, scored 38 runs from …     6                38      102
 2 Ali   England allrounder Test.1..Innings.2 Batting at number 6, scored 40 runs from …     6                40      64
 3 Ali   England allrounder Test.2..Innings.1 Batting at number 6, scored 25 runs from …     6                25      57
 4 Ali   England allrounder Test.2..Innings.2 Batting at number 7, scored 2 runs from 2…     7                2       20
 5 Ali   England allrounder Test.3..Innings.1 Batting at number 7, scored 0 runs from 2…     7                0       2
 6 Ali   England allrounder Test.3..Innings.2 Batting at number 7, scored 11 runs from …     7                11      56
```

Explanation:

Here we ensured the dataset is tame, categorical data (team, role and innings) are stored as factors, and numerical data (batting number, score and balls) are stored as integers as well as (batter and performance) are stored as character.

Cleaning Factor Levels:

R Code Input:

```r
# Load the forcats family (and dplyr)

library(dplyr)

library(forcats)


# Step 1: Inspect unique values to find typos

unique(ashes_long$team)

unique(ashes_long$role)


# Step 2: After checking the output, recode incorrect spellings

ashes_long <- ashes_long %>%

       mutate(

         team = fct_recode(team,

                           "England" = "English"),  # fix English → England

         role = fct_recode(role,

                           "allrounder" = "all rounder",

                           "allrounder" = "all-rounder",

                           "bowler" = "bowl",

                           "batter" = "bat",

                           "batter" = "batsman",

                           "batter" = "batting")

       )


# Step 3: Confirm cleaning worked

unique(ashes_long$team)

unique(ashes_long$role)
```

R Output:

```
# Step 3: Confirm cleaning worked
>    unique(ashes_long$team)
[1] England   Australia
Levels: Australia England
>    unique(ashes_long$role)
[1] allrounder   bowler      wicketkeeper batter
Levels: allrounder batter bowler wicketkeeper
```

Explanation:

We used fct_recode() to fix typos such as "English" → "England", "bowl" → "bowler".

This step ensures consistent factor levels so that later summaries and plots group correctly.

Summary

At this point:

- The dataset has 270 rows and is tidy.
- Each row represents one player's batting innings.
- All relevant values (team, role, score, balls) are in appropriate formats.

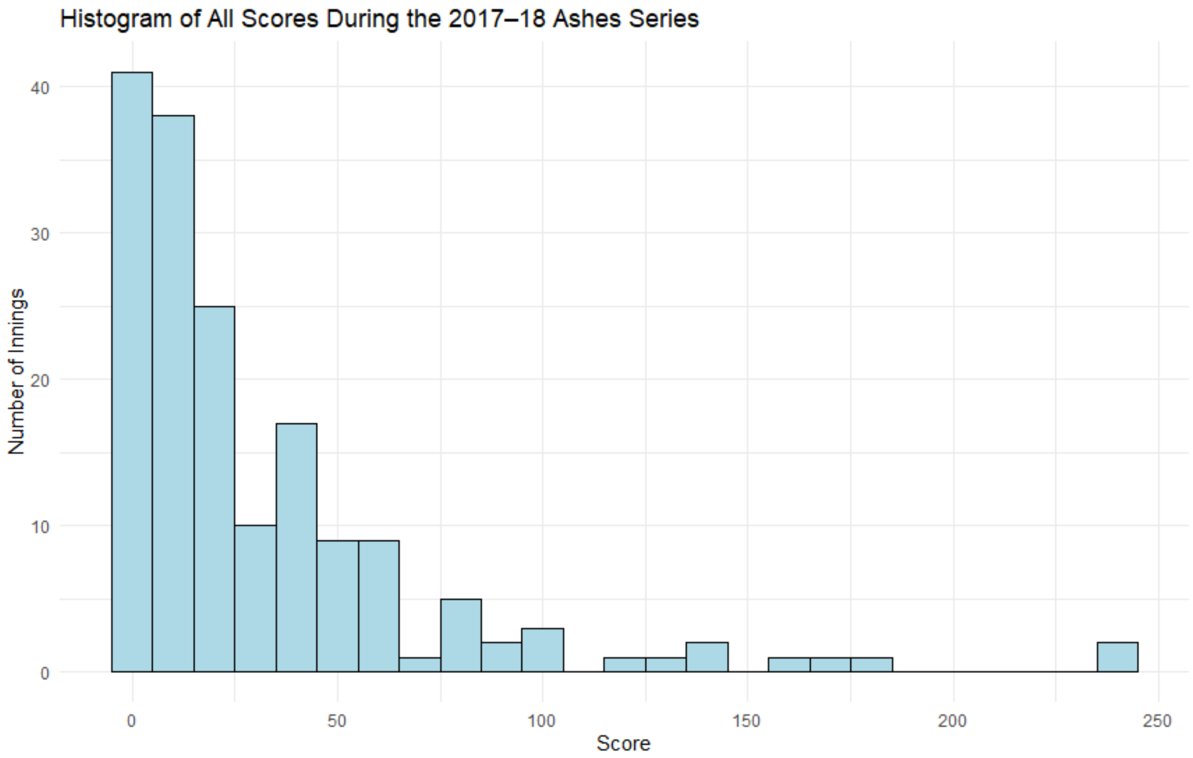
Univariate Analysis

Produce a histogram of all scores during the series.

R Code Input:

```
library(ggplot2)
     ggplot(ashes_long, aes(x = score)) +
       geom_histogram(binwidth = 10, fill = "lightblue", color = "black") +
       labs(
         title = "Histogram of All Scores During the 2017-18 Ashes Series",
         x = "Score",
         y = "Number of Innings"
       )
```

R Output:



Description of the Distribution of Scores:

The histogram shows a right-skewed distribution, with most player scores clustered at the lower end of the scale. The highest bars appear between 0 and 20 runs, indicating that low scores were the most common outcomes across all innings in the 2017–18 Ashes Series.

The location of the distribution, based on the summary statistics, is around 18 runs (median). Although the mean score is 32, the median provides a better measure of the typical performance because of the strong right skew. This suggests that most batters scored under 20 runs per innings, while a few high-scoring innings raised the mean.

The spread of the scores extends from 0 to over 200 runs, showing a wide range of batting performances from very low to exceptionally high scores.

A few outliers appear on the far right of the histogram, representing rare innings where players scored well above 150 or even 200 runs. These extreme values stretch the tail of the distribution to the right, confirming the skewed shape.

Calculating Number of Players per Team:

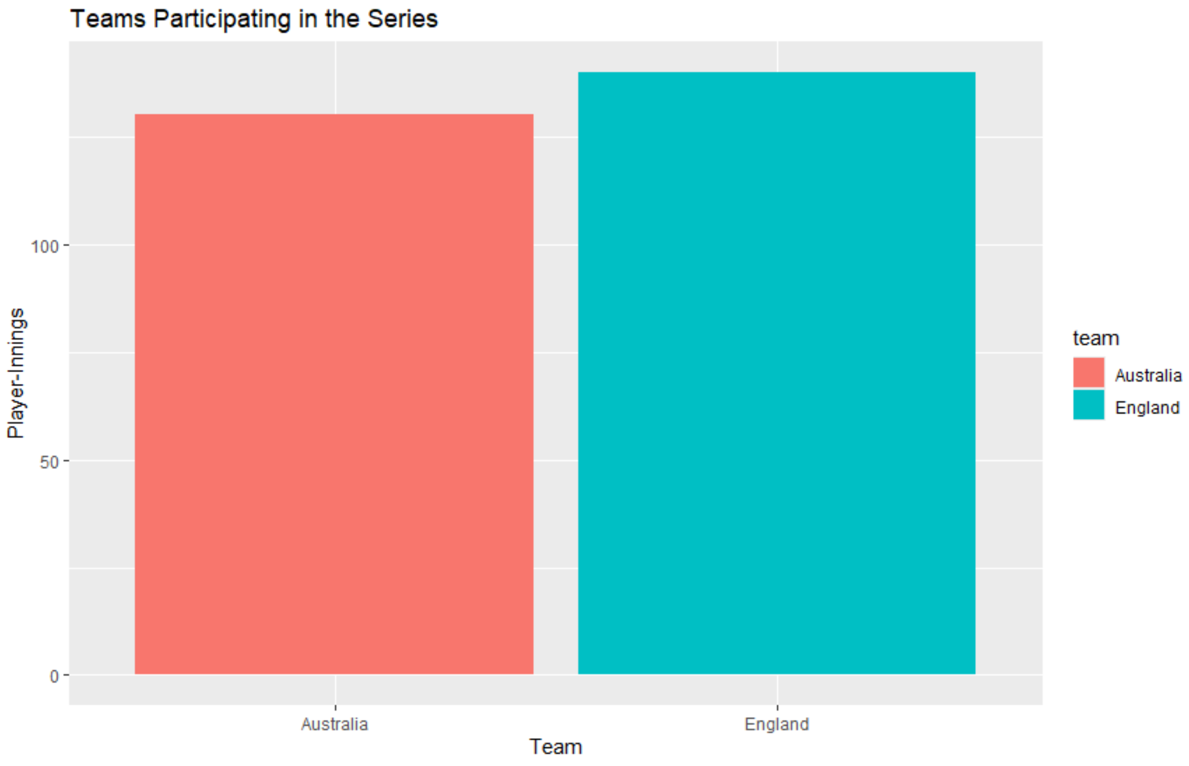
We will create a bar chart of the teams participating in the series.

R Code Input:

```
library(ggplot2)
library(dplyr)

# Step 1: Bar chart of teams
    ggplot(ashes_long, aes(x = team, fill = team)) +
      geom_bar() +
      labs(title = "Teams Participating in the Series",
           x = "Team",
           y = "Player-Innings")
 # Step 2: Count unique players per team
ashes_long %>%
      distinct(team, batter) %>%
      count(team)
```

R Output:



```
# A tibble: 2 × 2
  team        n
  <fct>     <int>
1 Australia    13
2 England      14
```

Explanation:

The bar chart shows the two teams that participated in the 2017–18 Ashes Series, Australia and England, each represented by a different colour. The height of each bar indicates the total number of player-innings recorded for each team. Because each player appears ten times in the dataset (two innings per Test across five Tests), this total reflects the number of players used by each team.

After counting the unique player names, Australia used 13 players, while England used 14 players. England used slightly more players than Australia, likely due to rotation or injuries throughout the series.

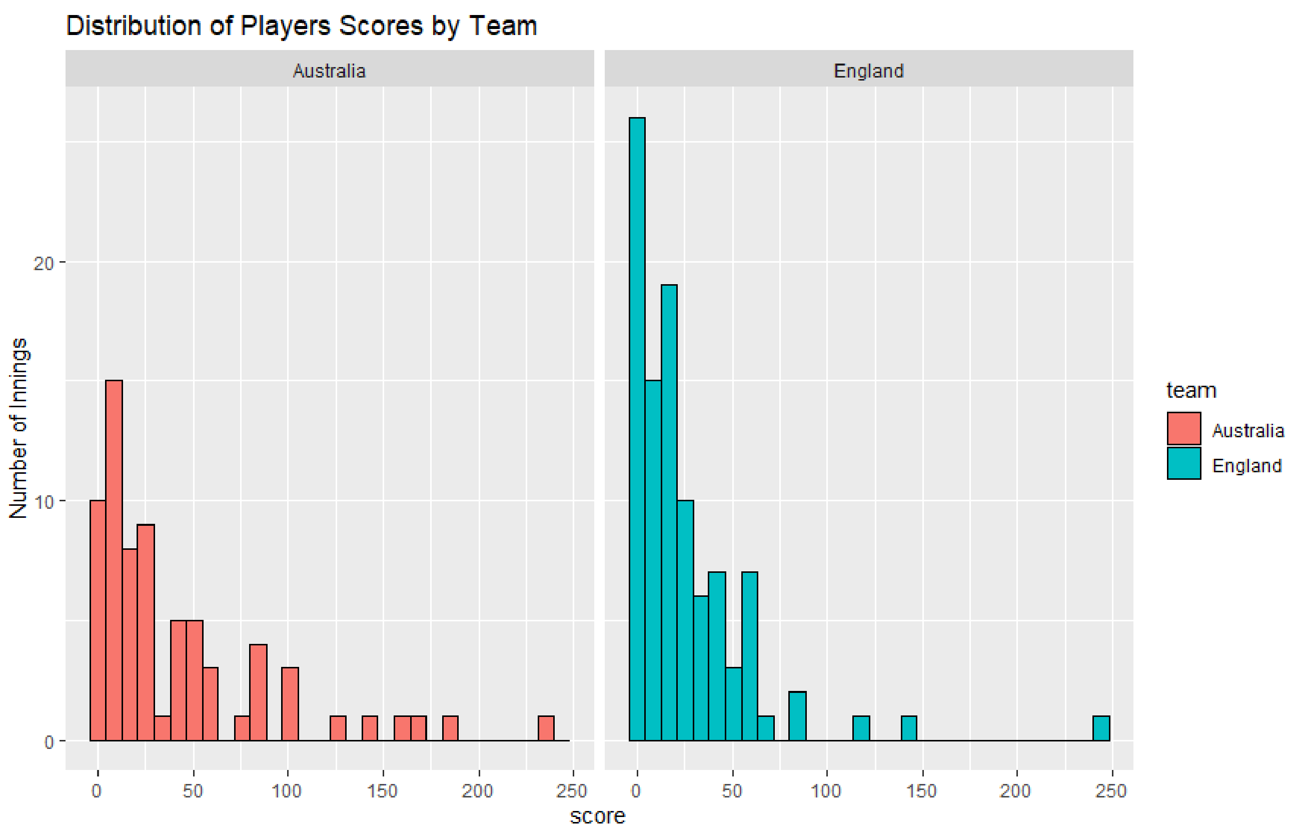
Scores for Each Team

Histograms of Scores Faceted by Team:

R Code Input:

```
ggplot(ashes_long, aes(x= score, fill= team)) +
      geom_histogram(color= "black") + facet_wrap(~team) +
      labs(title = "Distribution of Players Scores by Team",
           x= "score",
           y= "Number of Innings")
```

R Output:



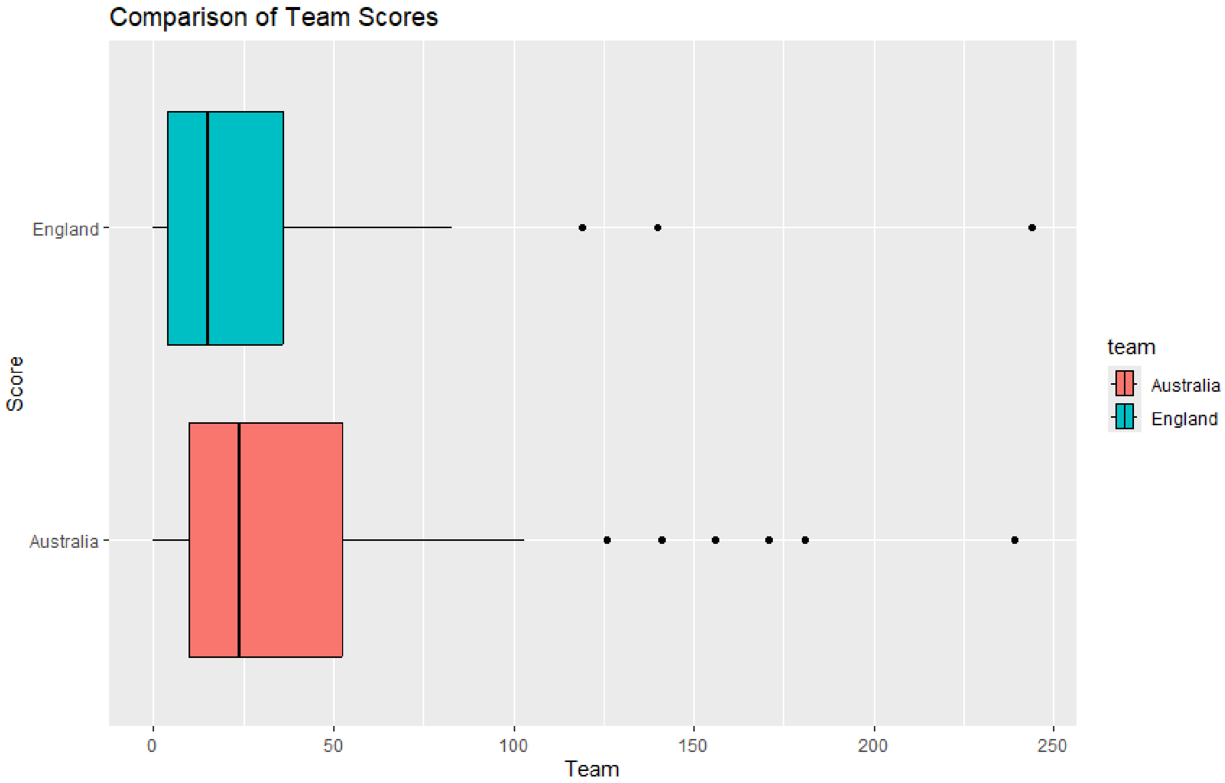
Explanation:
The faceted histograms show the distribution of player scores for each team in the 2017–18 Ashes Series. Each panel represents one team. Both teams display a similar right-skewed pattern, with most scores between 0 and 50 runs and a few very high innings extending beyond 150 runs. This indicates that low scores were common for both sides, while only a few players achieved exceptionally high totals.

Side-by-Side Boxplots of Scores by Team

R Code Input:

```
# Boxplots comparing scores by team
  ggplot(ashes_long, aes(x= score, y= team, fill= team)) +
      geom_boxplot(color= "black") +
      labs(title= "Comparison of Team Scores",
           x= "Team",
           y= "Score",
           )
```

R Output:



Explanation:

Boxplots summarise each team's score distribution. The line inside each box represents the median score, and the dots outside the whiskers show unusually high (or low) performances.

Comparison of Distributions:

Interpretation

- **Shape:** Both the histogram and boxplot show right-skewed distributions for both teams, with many low scores and a few exceptional high innings.
- **Location:** In the boxplot, the median line for Australia is slightly higher than for England, indicating stronger typical batting performances.

- **Spread:** The histogram shows that Australia’s scores are spread across a wider range, while the boxplot confirms more high outliers for Australia, suggesting a mix of low and very strong innings.
- **Outliers:** Both teams have several scores above 100, but Australia includes a few extreme outliers, possibly over 200 runs.

Conclusion:

Based on the boxplots and histograms, Australia appears to have achieved a higher average score and greater consistency in batting performances compared to England.

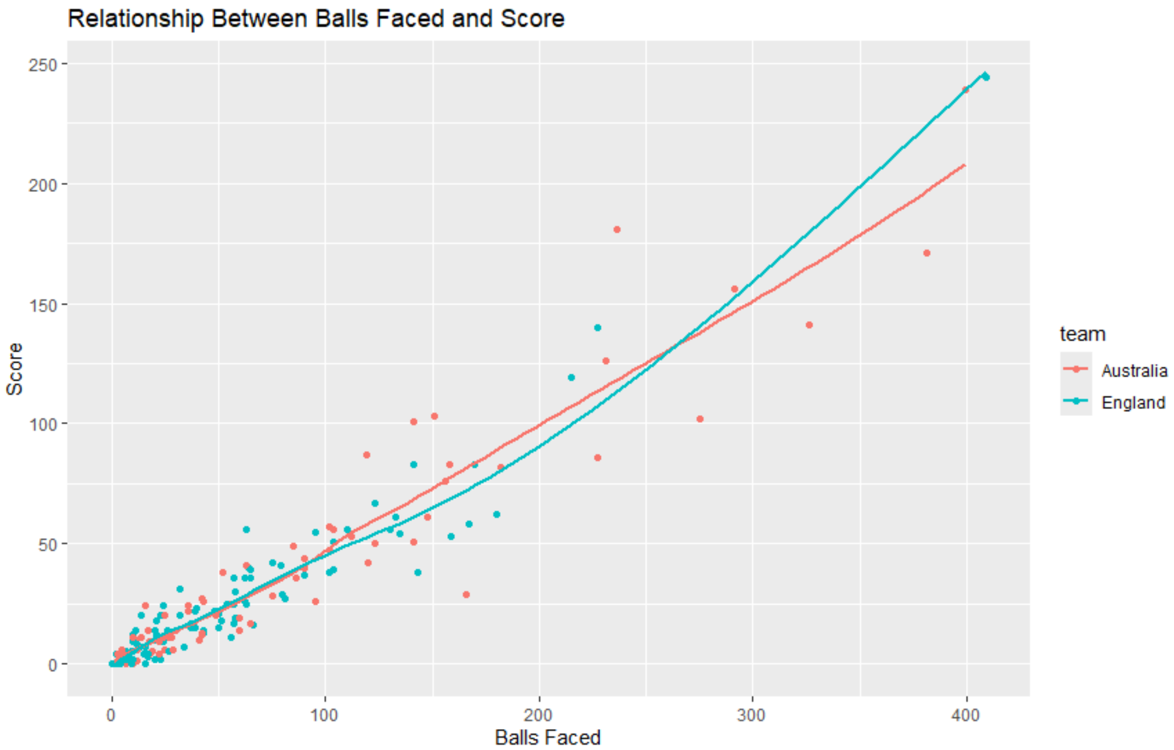
Scoring Rates

Scatterplot of Scores Against Number of Balls

R Code Input:

```
ggplot(ashes_long, aes(x = balls_faced, y = score, color = team)) +
      geom_point() +
      geom_smooth(se = FALSE) +
      labs(
        title = "Relationship Between Balls Faced and Score",
        x = "Balls Faced",
        y = "Score"
      )
```

R Output:


Relationship Between Balls Faced and Score

Explanation:

Each point on the scatterplot represents one batting innings. The upward trend shows a clear, logical relationship: batters who faced more balls generally scored more runs

<u>The Relationship between score and number of balls:</u>

Interpretation

- The scatterplot shows a positive, roughly linear, strong relationship between balls faced and runs scored.
- Points near the left and bottom represent innings with few balls and low scores.
- Points higher and to the right show longer innings with higher scores.
- A few points show players who faced many balls but scored modestly, likely due to defensive play.

Conclusion:

Yes, players who face more balls are generally more likely to score more runs.

Create a Scoring Rate Variable and Scatterplot

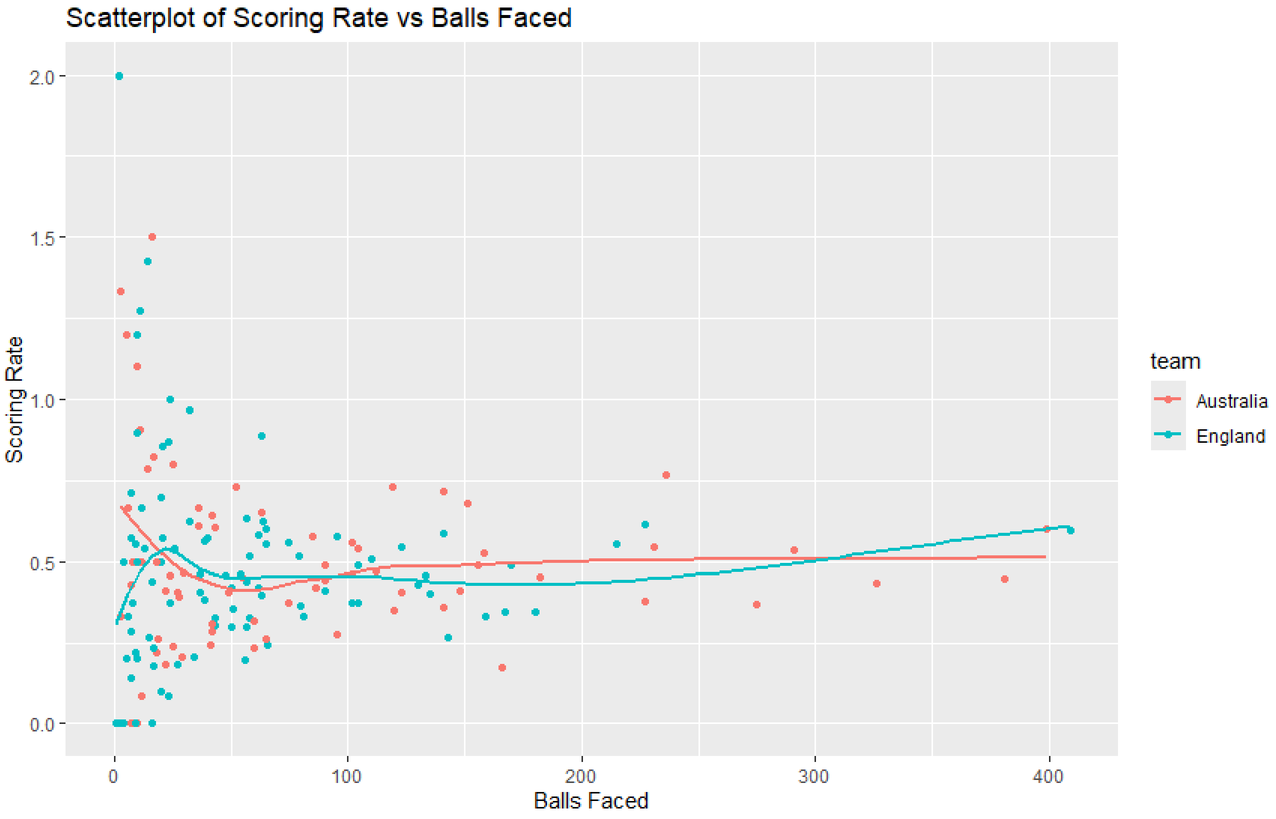
We will create a new variable, scoring rate, and produce a scatter plot of it against number of balls.

R Code Input:

```
# Create scoring_rate variable
ashes_long <- ashes_long %>%
      mutate(scoring_rate = score / balls_faced)

# Scatterplot of scoring rate vs balls
  ggplot(ashes_long, aes(x = balls_faced, y = scoring_rate, color = team)) +
        geom_point() +
        geom_smooth(se = FALSE) +
        labs(
          title = "Scatterplot of Scoring Rate vs Balls Faced",
          x = "Balls Faced",
          y = "Scoring Rate"
        )
```

R Output:



Explanation:

We computed a new variable — scoring_rate = runs / balls — to show how quickly a player scored. This gives insight into batting speed rather than total runs.

The Relationship Between Scoring Rate and Balls Faced:

Interpretation

- The scatterplot shows a weak negative relationship between scoring rate and balls faced.
- Players who faced only a few balls often have high or variable scoring rates (either quick runs or short, risky innings).
- As players face more balls, the scoring rate tends to stabilise or slightly decrease, suggesting that longer innings are usually steadier and less aggressive.

Conclusion

- There is no strong positive relationship between scoring rate and balls faced.
- Players who face more balls are not necessarily scoring faster; instead, they often play more cautiously to build longer innings.

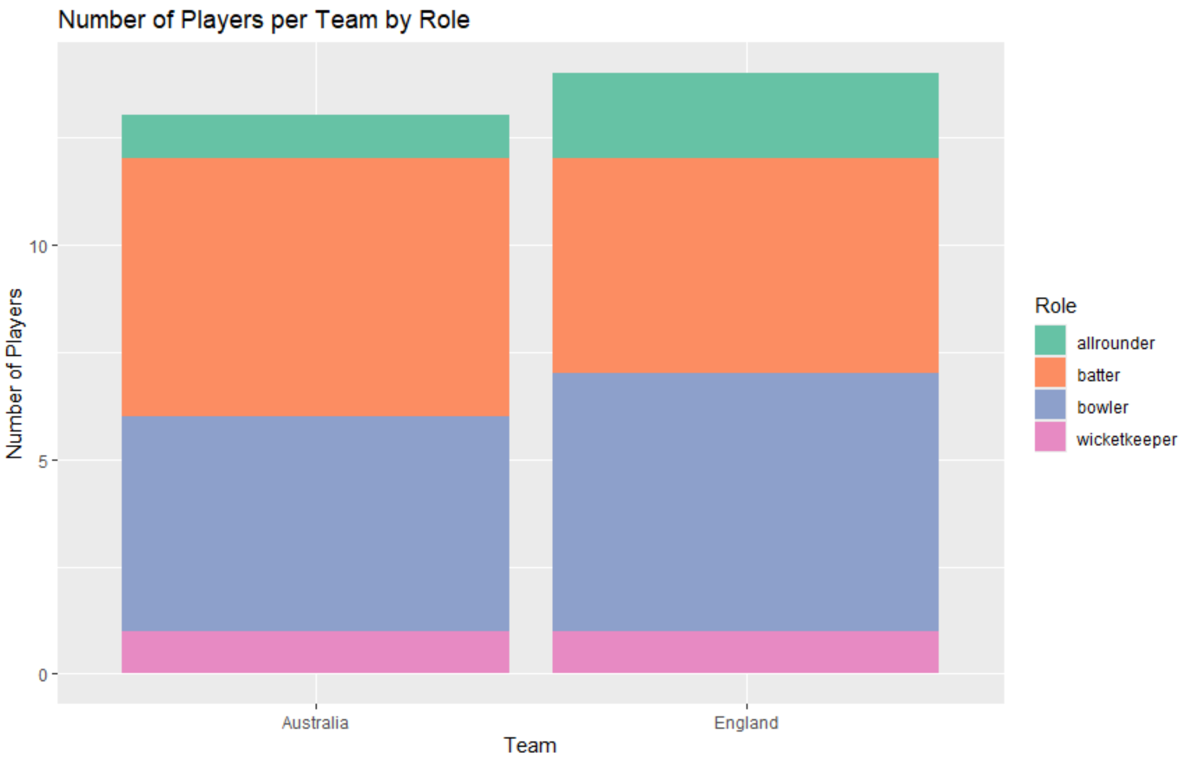
Teams' Roles:

Bar Chart of Player Roles per Team:

R Code Input:

```
# Bar chart of players per team coloured by their roles
  ggplot(ashes_long %>% distinct(batter, team, role),
         aes(x = team, fill = role)) +
    geom_bar(position = "stack") +
    labs(
      title = "Number of Players per Team by Role",
      x = "Team",
      y = "Number of Players",
      fill = "Role"
    )
```

R Output:



Explanation:

This chart visually compares how each team’s players are distributed by role. The height of each segment shows how many players of each role each team included.

Contingency Table of Proportions:

We will produce a contingency table of the proportion of players from each team who play in each particular role.

R Code Input:

```
# Proportion of players in each role by team (rounded to 2 decimal places)
      ashes_long %>%
        distinct(batter, team, role) %>%

        group_by(team, role) %>%

        summarise(count = n()) %>%

        mutate(proportion = round(count / sum(count), 2))
```

R Output:

```
 team        role          count  proportion
 <fct>       <fct>         <int>    <dbl>
1 Australia   allrounder      1       0.08
2 Australia   batter          6       0.46
3 Australia   bowler          5       0.38
4 Australia   wicketkeeper    1       0.08
5 England     allrounder      2       0.14
6 England     batter          5       0.36
7 England     bowler          6      0.43
8 England     wicketkeeper    1      0.07
```

Explanation:

The table shows the proportion of each team's players who are assigned to each role. It helps to understand team composition beyond just counts.

Interpretation

Using the bar chart and contingency table, it is clear that Australia is made up of a larger proportion of batters, while England contains a larger proportion of all-rounders.

From the data:

- Australia's team included 46% batters, compared to 36% for England.
- England had 14% all-rounders, compared to only 8% for Australia.

This indicates that Australia's lineup focused more on batting strength, relying on a higher number of specialist batters, whereas England's team relied more on flexibility, including a slightly greater share of all-rounders who could both bat and bowl.

Summary of Insights

Final Summary:

This analysis of the 2017–18 Men's Ashes series provides several insights into player performances and team composition.
Overall, the dataset revealed that batting scores across both teams followed a right-skewed distribution, where low scores were common, and a few innings reached very high totals. This suggests that consistency in batting performance was challenging across the series, with a few standout players making the difference.

When comparing teams, Australia displayed a slightly higher median score and a greater number of high outliers, suggesting stronger batting depth and more reliable top-order contributions. England, in contrast, showed more variation and fewer big innings, which may indicate instability in the batting lineup.

The relationship between balls faced and runs scored was strongly positive—players who stayed longer at the crease tended to score more. However, the scoring rate (runs per ball) slightly decreased as the number of balls faced increased, showing that long innings often involved slower scoring rates.

Finally, team composition analysis showed that Australia relied more heavily on specialist batters, while England employed more bowlers and all-rounders.
Together, these findings suggest that Australia's strategy favoured batting endurance, while England's team structure leaned toward bowling diversity.

References:

- University of Adelaide 2025, *R Data Visualisation and Analysis – Applied Data Science & Mathematics (MyUni)*, viewed 8 November 2025, https://myuni.adelaide.edu.au/courses/105626/modules.
- R Core Team 2024, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna. Available from: https://www.R-project.org/.
- Wickham, H & Grolemund, G 2017, *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, O'Reilly Media, Sebastopol. Available from: https://r4ds.hadley.nz.
- Wickham, H 2023, *The tidyverse package documentation*, R Studio PBC, Boston. Available from: https://www.tidyverse.org/packages/.
- Chang, W & Wickham, H 2024, *ggplot2: Elegant Graphics for Data Analysis – Reference Manual*, CRAN, viewed 8 November 2025, https://ggplot2.tidyverse.org/.
- Kuhn, M & Wickham, H 2023, *dplyr: A Grammar of Data Manipulation*, R Studio PBC, Boston. Available from: https://dplyr.tidyverse.org/.