Assessment 1 – Data set

A1844790 - Dylan Rohan

Assignment completed using R and RStudio(R Core Team 2021). Loading libraries (tidyverse, stringr, dplyr, and gmodel) and uploading/checking csv file data (Warnes et al. 2018; Wickham 2019; Wickham et al. 2019; Wickham et al. 2021).

```
library(tidyverse)
library(dplyr)
library(stringr)
library(rmarkdown)

ashes <- read_csv("C:\\Users\\rohad\\OneDrive\\Documents\\Data science\\Data Taming, modelling and Vizalization_RStudio\\a1\\a1\\ashes.csv")
#double slashes for windows directory
knitr::kable(head(ashes), caption = "Table 1: Uploaded ashes data, currently a tibble of 27 x 13.")
```

Table 1: Uploaded ashes data, currently a table of 27 x 13.

batter	team	role	Test 1, Innings 1	Test 1, Innings 2	Test 2, Innings 1	Test 2, Innings 2	Test 3, Innings 1	Test 3, Innings 2	Test 4, Innings 1	Test 4, Innings 2	Test 5, Innings 1	Test 5, Innings 2
Ali	England	allrounder	Batting at number 6, scored 38 runs from 102 balls including 2 fours and 1 sixes.	Batting at number 6, scored 40 runs from 64 balls including 6 fours and 0 sixes.	Batting at number 6, scored 25 runs from 57 balls including 2 fours and 0 sixes.	Batting at number 7, scored 2 runs from 20 balls including 0 fours and 0 sixes.	Batting at number 7, scored 0 runs from 2 balls including 0 fours and 0 sixes.	Batting at number 7, scored 11 runs from 56 balls including 2 fours and 0 sixes.	Batting at number 7, scored 20 runs from 14 balls including 2 fours and 1 sixes.	Batting at number NA, scored NA including NA fours and NA sixes.	Batting at number 7, scored 30 runs from 58 balls including 2 fours and 0 sixes.	Batting at number 7, scored 13 runs from 43 balls including 1 fours and 0 sixes.
Anderson	English	bowl	Batting at number 11, scored 5 runs from 9 balls including 1 fours and 0 sixes.	Batting at number 11, scored 0 runs from 1 balls including 0 fours and 0 sixes.	Batting at number 11, scored 0 runs from 3 balls including 0 fours and 0 sixes.	Batting at number 11, scored 0 runs from 0 balls including 0 fours and 0 sixes.	Batting at number 11, scored 0 runs from 7 balls including 0 fours and 0 sixes.	Batting at number 11, scored 1 runs from 7 balls including 0 fours and 0 sixes.	Batting at number 11, scored 0 runs from 16 balls including 0 fours and 0 sixes.	Batting at number NA, scored NA including NA fours and NA sixes.	Batting at number 11, scored 0 runs from 3 balls including 0 fours and 0 sixes.	Batting at number 11, scored 2 runs from 23 balls including 0 fours and 0 sixes.
Bairstow	England	wicketkeeper	Batting at number 7, scored 9 runs from 24 balls including 1 fours and 0 sixes.	Batting at number 7, scored 42 runs from 75 balls including 2 fours and 1 sixes.	Batting at number 7, scored 21 runs from 50 balls including 2 fours and 0 sixes.	Batting at number 8, scored 36 runs from 57 balls including 5 fours and 0 sixes.	Batting at number 6, scored 119 runs from 215 balls including 18 fours and 0 sixes.	Batting at number 6, scored 14 runs from 26 balls including 3 fours and 0 sixes.	Batting at number 6, scored 22 runs from 39 balls including 3 fours and 0 sixes.	Batting at number NA, scored NA including NA fours and NA sixes.	Batting at number 6, scored 5 runs from 7 balls including 1 fours and 0 sixes.	Batting at number 6, scored 38 runs from 143 balls including 4 fours and 0 sixes.
Ball	England	bowl	Batting at number 10, scored 14 runs from 11 balls including 3 fours and 0 sixes.	Batting at number 10, scored 1 runs from 5 balls including 0 fours and 0 sixes.	Batting at number NA, scored NA including NA fours and NA sixes.	Batting at number NA, scored NA including NA fours and NA sixes.	Batting at number NA, scored NA including NA fours and NA sixes.	Batting at number NA, scored NA including NA fours and NA sixes.	Batting at number NA, scored NA including NA fours and NA sixes.	Batting at number NA, scored NA including NA fours and NA sixes.	Batting at number NA, scored NA including NA fours and NA sixes.	Batting at number NA, scored NA including NA fours and NA sixes.
Bancroft	Australia	bat	Batting at number 1, scored 5 runs from 19 balls including 0 fours and 0 sixes.	Batting at number 1, scored 82 runs from 182 balls including 10 fours and 1 sixes.	Batting at number 1, scored 10 runs from 41 balls including 0 fours and 0 sixes.	Batting at number 1, scored 4 runs from 8 balls including 1 fours and 0 sixes.	Batting at number 1, scored 25 runs from 55 balls including 3 fours and 0 sixes.	Batting at number NA, scored NA including NA fours and NA sixes.	Batting at number 1, scored 26 runs from 95 balls including 2 fours and 0 sixes.	Batting at number 1, scored 27 runs from 42 balls including 4 fours and 0 sixes.	Batting at number 1, scored 0 runs from 7 balls including 0 fours and 0 sixes.	Batting at number NA, scored NA including NA fours and NA sixes.
Bird	Australia	bowl	Batting at number NA, scored NA including NA fours and NA sixes.	Batting at number NA, scored NA including NA fours and NA sixes.	Batting at number NA, scored NA including NA fours and NA sixes.	Batting at number NA, scored NA including NA fours and NA sixes.	Batting at number NA, scored NA including NA fours and NA sixes.	Batting at number NA, scored NA including NA fours and NA sixes.	Batting at number 9, scored 4 runs from 6 balls including 1 fours and 0 sixes.	Batting at number NA, scored NA including NA fours and NA sixes.	Batting at number NA, scored NA including NA fours and NA sixes.	Batting at number NA, scored NA including NA fours and NA sixes.

```
unique(ashes$team)
## [1] "England"   "English"   "Australia"
#need to correct variable English to be England
unique(ashes$role)
## [1] "allrounder"   "bowl"         "wicketkeeper" "bat"            "bowler"
## [6] "batting"      "batsman"      "all rounder"  "all-rounder"
#many duplicates under alternate variable names, eg. bat, batsman, batting
```

Question One: Reading and Cleaning

1.1

For our analysis, the subjects are not the cricketers themselves, but each batting innings they participated in. In order to make the data tidy **each subject needs its own row**. Rearrange the data into a long format so that there is a row for each batter in each innings. Your new tibble should have 270 rows. [2 points]

Each cell should represent only one measurement. Use str_match() to create new columns for each of the following for each player innings:

- **The batting order, their score, & the number of balls they faced**. [2 points]

```
colnames(ashes)
#Getting the column titles for conversion
ashes_longform <- gather(ashes, key = "innings", value = "description", "Test 1, Innings 1" : "Test 5, Innings 2")
#tibble now in long form, 270 x 5
ashes_innings_first <- ashes_longform[c(4, 1, 2, 3, 5)]
#subject first
knitr::kable(head(ashes_innings_first), caption = "Table 2: Table now in long form with subject first.")
```

Table 2: Table now in long form with subject first.

innings	batter	team	role	description
Test 1, Innings 1	Ali	England	allrounder	Batting at number 6, scored 38 runs from 102 balls including 2 fours and 1 sixes.
Test 1, Innings 1	Anderson	English	bowl	Batting at number 11, scored 5 runs from 9 balls including 1 fours and 0 sixes.
Test 1, Innings 1	Bairstow	England	wicketkeeper	Batting at number 7, scored 9 runs from 24 balls including 1 fours and 0 sixes.
Test 1, Innings 1	Ball	England	bowl	Batting at number 10, scored 14 runs from 11 balls including 3 fours and 0 sixes.
Test 1, Innings 1	Bancroft	Australia	bat	Batting at number 1, scored 5 runs from 19 balls including 0 fours and 0 sixes.
Test 1, Innings 1	Bird	Australia	bowl	Batting at number NA, scored NA including NA fours and NA sixes.

```
order <- str_match(ashes_innings_first$description, "Batting at number ..")
with_order <- cbind(ashes_innings_first, order)
#Order now has its own column, values are strings
runs <- str_match(with_order$description, "scored ....")
with_runs <- cbind(with_order, runs)
#runs now has its own column, values are strings
no._of_balls <- str_match(with_runs$description, "from ....")
all_columns<- cbind(with_runs, no._of_balls)
#no.  of balls now has its own column, values are
batting_order <- str_replace_all(all_columns$order, "[^0-9.-]", "")
runs_ <- str_replace_all(all_columns$runs, "[^0-9.-]", "")
```

```
balls_ <- str_replace_all(all_columns$no._of_balls, "[^0-9.-]", "")
#Taking numerical values from strings
order1 <- tibble(batting_order)
runs1 <- tibble(runs_)
balls1 <- tibble(balls_)
#making data frames from those values
a1_o <- cbind(ashes_innings_first, order1)
a1_o_r <- cbind(a1_o, runs1)
a1_o_r_b<- cbind(a1_o_r, balls1)
#Order same, so binding columns
a1o_r_b <- a1_o_r_b$description <- NULL
a1_o_r_b <- a1_o_r_b %>%
  mutate_all(na_if, "")
#removing description column
knitr::kable(head(a1_o_r_b), caption = "Table 3: Now a table of 270 x 7 (remo
ved description, but it's still accessible in 'ashes_innings_first'.")
```

Table 3: Now a table of 270 x 7 with batting order, run, and ball attributes (removed description, but it's still accessible in 'ashes_innings_first').

innings	batter	team	role	batting_order	runs_	balls_
Test 1, Innings 1	Ali	England	allrounder	6	38	102
Test 1, Innings 1	Anderson	English	bowl	11	5	9
Test 1, Innings 1	Bairstow	England	wicketkeeper	7	9	24
Test 1, Innings 1	Ball	England	bowl	10	14	11
Test 1, Innings 1	Bancroft	Australia	bat	1	5	19
Test 1, Innings 1	Bird	Australia	bowl	NA	NA	NA

```
#__________Alternative method_____________________________#

trial <- ashes_innings_first %>%
  mutate("runs_"=str_match(description,"from ....") , "batting_order" = str_m
atch(description, "Batting at number .."), "balls_" = str_match(description,
"scored ...."))
#description string broken into appropriate columns
trial <- trial %>%
  mutate("runs_" = str_replace_all(trial$runs_, "[^0-9.-]",""), "balls_"=str_
replace_all(trial$balls_, "[^0-9.-]",""), "batting_order"=str_replace_all(tri
al$batting_order, "[^0-9.-]",""))
trial <- mutate_all(trial, na_if, "")
#Left the description column in here, but all is right with the world
knitr::kable(head(trial), caption = "Table 4: Alternate method to get the sam
e answer.")
```

Table 4: Alternate method to get the same answer.

innings	batter	team	role	description	runs_	batting_order	balls_
Test 1, Innings 1	Ali	England	allrounder	Batting at number 6, scored 38 runs from 102 balls including 2 fours and 1 sixes.	102	6	38
Test 1, Innings 1	Anderson	English	bowl	Batting at number 11, scored 5 runs from 9 balls including 1 fours and 0 sixes.	9	11	5
Test 1, Innings 1	Bairstow	England	wicketkeeper	Batting at number 7, scored 9 runs from 24 balls including 1 fours and 0 sixes.	24	7	9
Test 1, Innings 1	Ball	England	bowl	Batting at number 10, scored 14 runs from 11 balls including 3 fours and 0 sixes.	11	10	14
Test 1, Innings 1	Bancroft	Australia	bat	Batting at number 1, scored 5 runs from 19 balls including 0 fours and 0 sixes.	19	1	5
Test 1, Innings 1	Bird	Australia	bowl	Batting at number NA, scored NA including NA fours and NA sixes.	NA	NA	NA

```
# ____________________________________________________________#
```

1.2

Recode the data to make it 'tame', that is:

- Ensure all categorical variables with a small number of levels are coded as factors
    - Innings (10 ordered levels), team (two levels), role (three levels), & batting order (10 ordered levels)
- Ensure all categorical variables with a large number of levels are coded as characters,
    - Player (the amount of players they had to choose from is very large or even unknown and can change from series to series).
- Ensure all quantitative variables are coded as integers or numeric, as appropriate. [3 points]
    - Runs & balls – discrete numerical values.

```
ashes_tibble <- as_tibble(a1_o_r_b)
#making a data frame from a1_o_r_b to set value type
ashes_tibble$batting_order <- as.factor(ashes_tibble$batting_order)
#low level ordinal, label = factor
ashes_tibble$runs_ <- as.integer(ashes_tibble$runs_)
ashes_tibble$balls_ <- as.integer(ashes_tibble$balls_)
#countable, discrete = integer
ashes_tibble$innings <- as.factor(ashes_tibble$innings)
#innings total = 10 (low level and ordinal), is a label/name = factors
ashes_tibble <- rename(ashes_tibble,"player"="batter")
ashes_tibble$player <- as.character(ashes_tibble$player)
#player makes more sense as a variable name. The teams have several people th
at could take the position, categorical variable of unknown levels = characte
r.
ashes_tibble$team <- as.factor(ashes_tibble$team)
ashes_tibble$role <- as.factor(ashes_tibble$role)
```

```
#both low value labels, so factors
#demonstrating the value types have been set:
ashes_tibble
```

```
## # A tibble: 270 x 7
##    innings           player   team      role         batting_order runs_ balls_
##    <fct>             <chr>    <fct>     <fct>        <fct>         <int>  <int>
```

1.3

Clean the data; recode the factors using fct_recode() such that there are no typographical errors in the team names and player roles. [2 points]

```
unique(ashes_tibble$player)
summary(unique(ashes_tibble$innings))
unique(ashes_tibble$team)
## [1] England   English   Australia
## Levels: Australia England English
unique(ashes_tibble$role)
## [1] allrounder   bowl         wicketkeeper bat          bowler
## [6] batting      batsman      all rounder  all-rounder
## 9 Levels: all-rounder all rounder allrounder bat batsman batting ... wicke
tkeeper
#English to England, and unify player roles
ashes_corrected_ <- ashes_tibble %>%
  mutate(team = fct_recode(team, "England" = "English"))%>%
  mutate(role = fct_recode(role, "all-rounder" = "allrounder", "all-rounder"=
"all rounder", "batsman"="batting", "batsman"="bat", "bowler"="bowl"))
ac <- ashes_corrected_
knitr::kable(head(ac), caption = "Table 5: Table demonstrating the data is no
w clean and tame")
```

Table 5: Table demonstrating the data is now clean and tame

innings	player	team	role	batting_order	runs_	balls_
Test 1, Innings 1	Ali	England	all-rounder	6	38	102
Test 1, Innings 1	Anderson	England	bowler	11	5	9
Test 1, Innings 1	Bairstow	England	wicketkeeper	7	9	24
Test 1, Innings 1	Ball	England	bowler	10	14	11
Test 1, Innings 1	Bancroft	Australia	batsman	1	5	19
Test 1, Innings 1	Bird	Australia	bowler	NA	NA	NA

Question two: univariate analysis

2.1

Produce a histogram of all scores during the series. [1 point]

```
#Histogram default below, bin of 30
ggplot(ac)+geom_histogram(aes(x=runs_, ), fill= "black", na.rm=TRUE) +
  ggtitle("The Runs Achieved Over An Innings in the \n2017/18 Ashes Series")+
  labs(x= "Scores reached", y ="Frequency")
```

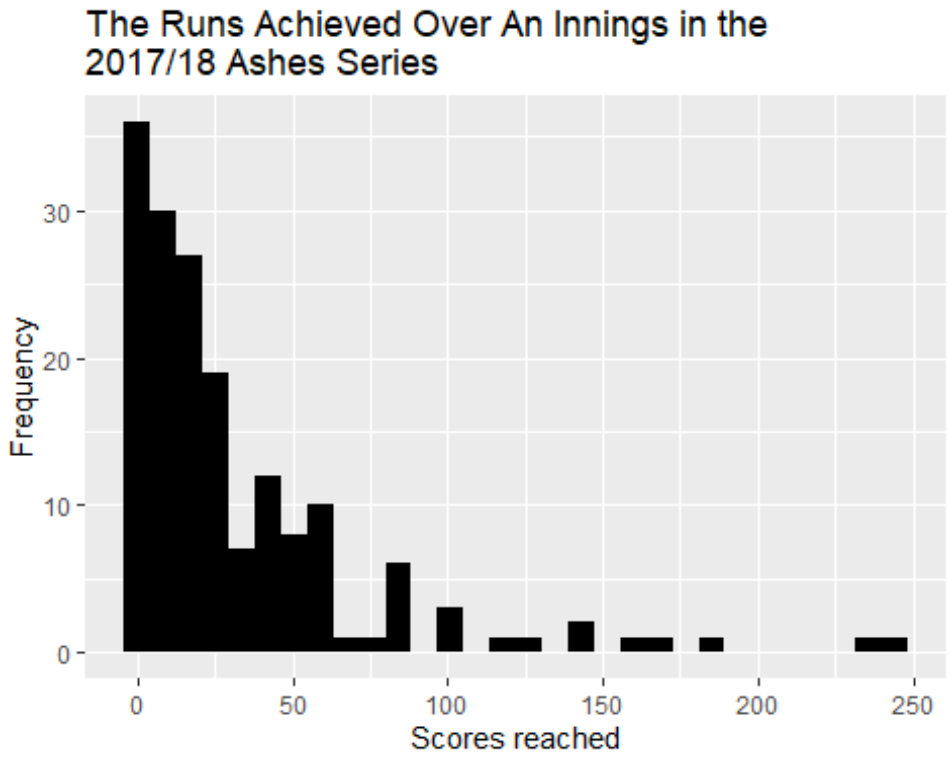

Figure 1: Histogram showing the frequency of scores reached in individual batting performances across the 2017/18 ashes series.

```
#ac$runs_ %>%
#  unique()
#cool find -> 70 unique values excluding NA, bin of 70 width = 1 for a bar ch
art as below
#ggplot(ac)+geom_histogram(mapping = aes(x=runs_), na.rm=TRUE, bins=70, binwi
dth = 1)+
#  ggtitle("Total runs acheieved")+labs(x= "Total runs")
```

2.2

Describe the distribution of scores, considering shape, location spread and outliers. [4 points]

```
summary(ac$runs_, na.rm = TRUE)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    6.00   18.00   32.09   41.00  244.00     101

range(ac$runs_, na.rm = TRUE, finite= TRUE)

## [1]   0 244

sd(ac$runs_, na.rm = TRUE)

## [1] 41.30805

table(ac$runs_)

##
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
##  12   4   8   3   9   5   4   3   2   4   3   6   3   2   8   3   1   3   2   2
##  20  21  22  23  24  25  26  27  28  29  30  31  36  37  38  39  40  41  42  44
##   5   1   3   1   3   4   3   2   1   2   1   1   4   1   3   2   2   2   2   1
##  47  49  50  51  53  54  55  56  57  58  61  62  67  76  82  83  86  87 101 102
##   1   1   1   2   2   1   1   4   1   1   2   1   1   1   1   3   1   1   1   1
## 103 119 126 140 141 156 171 181 239 244
##   1   1   1   1   1   1   1   1   1   1
```

This is a right-skewed shaped graph (*figure 1*). The mean score was 32 with a standard deviation of 41. Two players achieved scores over 200 which pulled the mean away from the mode of 12, and median of 18. The interquartile range was 35, the domain was [0,244], and the range was [0,25]. An outlier is anything 1.5 x interquartile range (IQR) from the edges of the IQR. Functionally, this indicates that a score higher than (1.5x35+41=) 94 is an outlier. With that definition, there were 11 outliers over the series.

2.3

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? [3 points]

Australia had 13 players, while England had 14.

```
ggplot(ac, aes(x= runs_, col=team))+geom_bar()
#^this maps every players innings, we need to combine player scores across th
e innings

indiv_runs <-  ac%>%
  group_by(player) %>%
  summarise(team,role,runs_in_series = sum(runs_, na.rm=TRUE))%>%
  unique()
```

```
knitr::kable(head(indiv_runs), caption = "Table 6: Demonstrating the scores h
ave been totaled for each player")
```

Table 6: Demonstrating the scores have been totaled for each player so that each player is represented by a single row.

player	team	role	runs_in_series
Ali	England	all-rounder	179
Anderson	England	bowler	8
Bairstow	England	wicketkeeper	306
Ball	England	bowler	15
Bancroft	Australia	batsman	179
Bird	Australia	bowler	4

```
unique(ac$player)
#all players accounted for

ggplot(indiv_runs, aes(x=team, fill=team))+
  geom_bar()+ggtitle("Number of Players On Each Team in the \n2017/18 Ashes S
eries")+
  scale_y_continuous(breaks = seq(0, 20, by = 1))+
  labs(x = "Team", y= "Number of players")
```

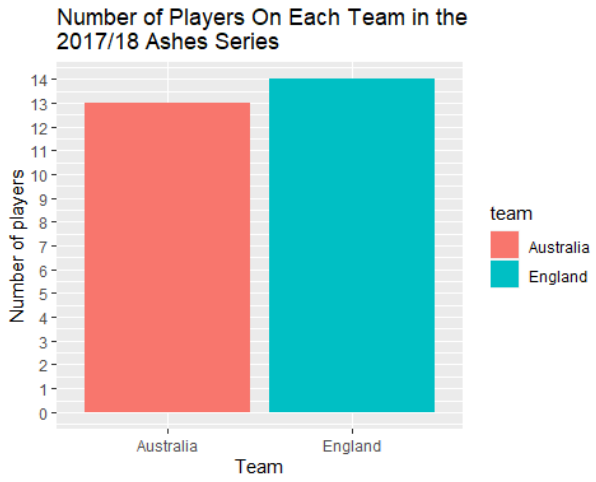

Figure 2: Bar chart indicating the number of players on each team during the 2017/18 Ashes series.

```
#players per team^

          #_____________________________________________________
#What I thought question 2.3 wanted
indiv_runs %>%
  ggplot(aes(x=player, y=runs_in_series, fill=role))+
  geom_bar(stat="identity")+
```

```
  ggtitle("Individual performance over the \n2017/18 Ashes series")+
  labs(x = "", y= "")+
  theme(axis.text.x= element_text(angle =-90, hjust = 0))
```

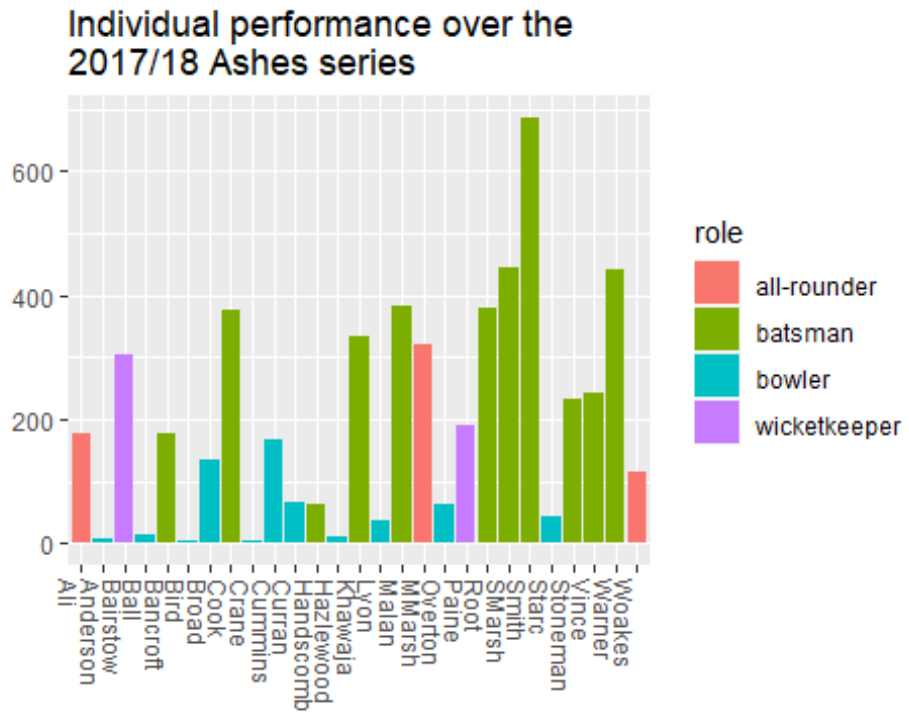

Figure 3: A bonus bar chart of the individual player performances over the 2017/18 Ashes series with colour indicating their role in the team.

```
#score per player
          #_____________________________________________________________
```

Question Three: Scores for each team

3.1

Using ggplot, produce histograms of scores during the series, faceted by team. [1 point]

```
ac %>%
  ggplot(aes(x=runs_, fill=team))+
  geom_histogram(show.legend = FALSE)+
  scale_y_continuous(breaks = seq(0, 30, by = 1))+
  facet_wrap(~team)+
  ggtitle("Team Batting Performance in the \n2017/18 Ashes Series")+
  labs(x = "Score", y= "Frequency")
```

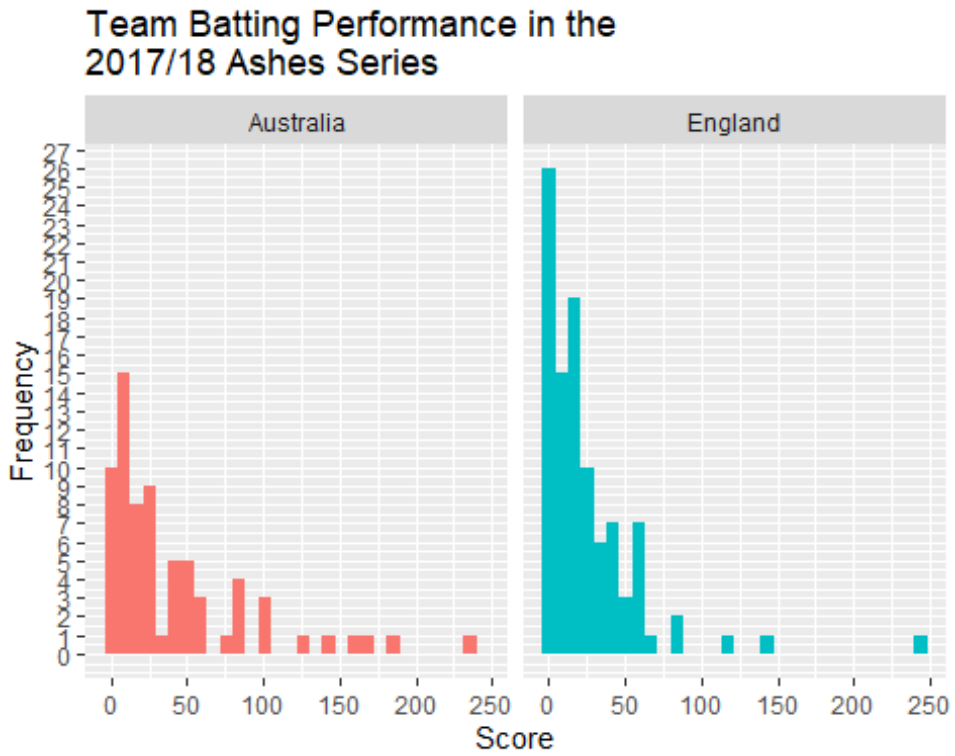

Figure 4: Faceted histograms indicating the frequency of scores reached for each team during the 2017/18 Ashes series; bin=30.

3.2

Produce side-by-side boxplots of scores by each team during the series. [1 point]

(Side by side as in facet grid? If it's just the normal output you're after see figure 6).

```
ac %>%
  ggplot(aes(y=runs_, fill=team))+
  geom_boxplot(show.legend = FALSE)+
  facet_grid(~team)+
  ggtitle("Boxplot of Team Batting Performance over the \n2017/18 Ashes Series")+
  labs(x = "Team", y="Runs over the series")
```

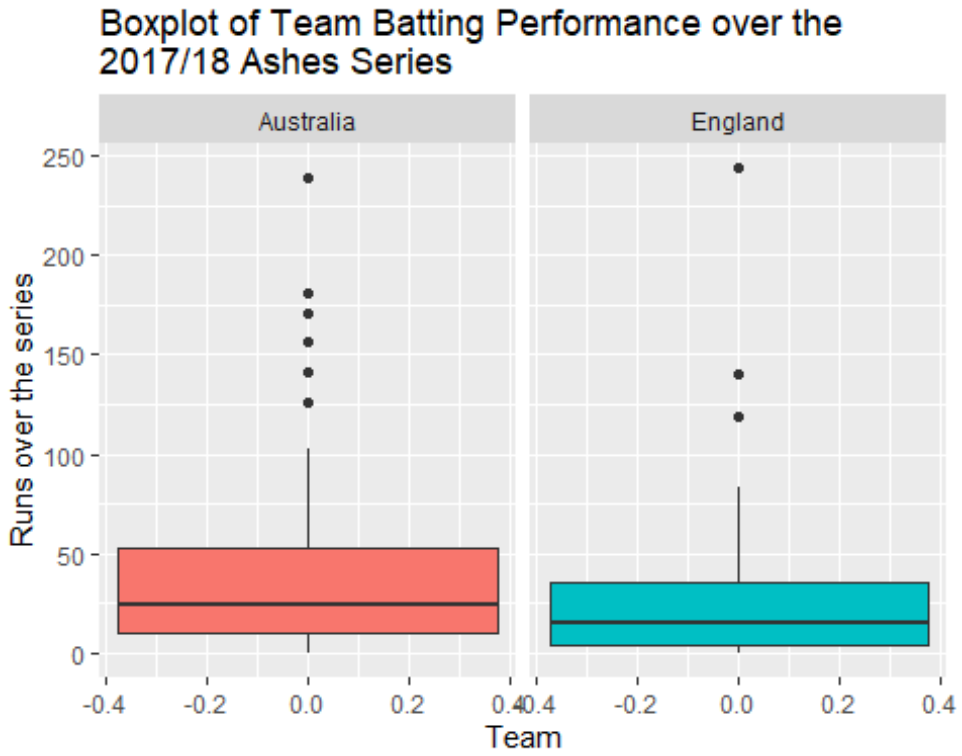

Figure 5: Boxplot representing the runs reached per batsman in an innings over the 2017/18 Ashes series for each team.

3.3

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 average score? [5 points]

```
#ENGLISH INDIVIDUALS
england_players <- ac[ac$team != "Australia", ]
summary(england_players$runs_, na.rm =TRUE)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    4.00   15.00   25.28   36.00  244.00      41

sd(england_players$runs_, na.rm = TRUE)

## [1] 33.61336

#England's statistics


#AUSTRALIAN INDIVIDUALS
aus_players <- ac[ac$team != "England",]
summary(aus_players$runs_, na.rm= TRUE)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   10.00   24.00   41.71   52.50  239.00      60

sd(aus_players$runs_, na.rm = TRUE)

## [1] 48.88174

#for outliers
ggplot(ac, aes(x = team, y = runs_, fill =team)) +
  geom_boxplot(show.legend = FALSE) +
  stat_summary(
    aes(label = round(stat(y), 1)),
    geom = "text",
    fun.y = function(y) { o <- boxplot.stats(y)$out; if(length(o) == 0) NA else o },
    hjust = -1)+
  ggtitle ("Boxplot of Team Batting Performance over the \n2017/18 Ashes Series")+
  labs(x = "Team", y="Runs over the series")
```

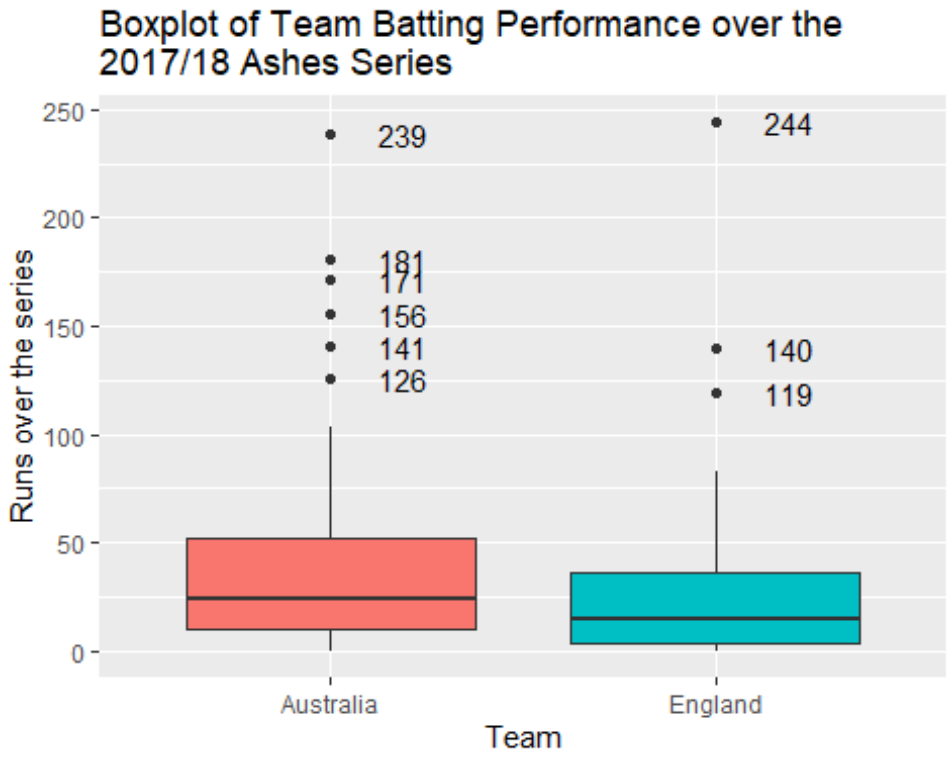

Figure 6: Boxplots representing the spread of runs reached by players in each team over the 2017/18 Ashes seies with outliers labelled with their values.

Both teams have right-skewed histograms, indicating higher scores are less common than lower scores (*figure 4*). England was more right-skewed than Australia due to it having a greater proportion of players reaching lower scores. Using the default bin width of 30, the domains are similar for both teams, [0,~250). But ranges differed, England had more players end the innings with scores lower than 50 so their range is much larger, [0, 26]; Australia's range is lower, sitting at [0, 15]. Australia appears to have had the highest average score. The mean score total was located at 42 for Australia with standard deviation of 49 (median of 24, and bimodal; 4 and 11), but only 25 for England with a standard deviation of 34 (median of 15, mode of 2). The spread also differed, the IQR was 32 for England, and 43 for Australia. This indicates that the English performed more consistently, around a lower mean score while Australia's scores varied more, with a few higher scores that pulled the mean higher. According to the boxplots (*figure 5 and 6*), there were six outliers for Australia (126, 141, 156, 171, 181, and 239), and three for England (119, 140, 244). That's six players that reached a score above (1.5 x IQR + 3rd quartile) 116 for Australia, and three above 84 for England.

Question Four: Scoring rates

4.1

Produce a scatterplot of scores against number of balls. [1 point]

```
ggplot(ac, aes( x = runs_, y= balls_, col=team))+
  geom_point()+
  geom_smooth()+
  ggtitle("Relationship between Balls Faced and \nScore Reached in the 2017/1
8 Ashes Series")+
  labs(x = "Score reached", y="Balls")
```

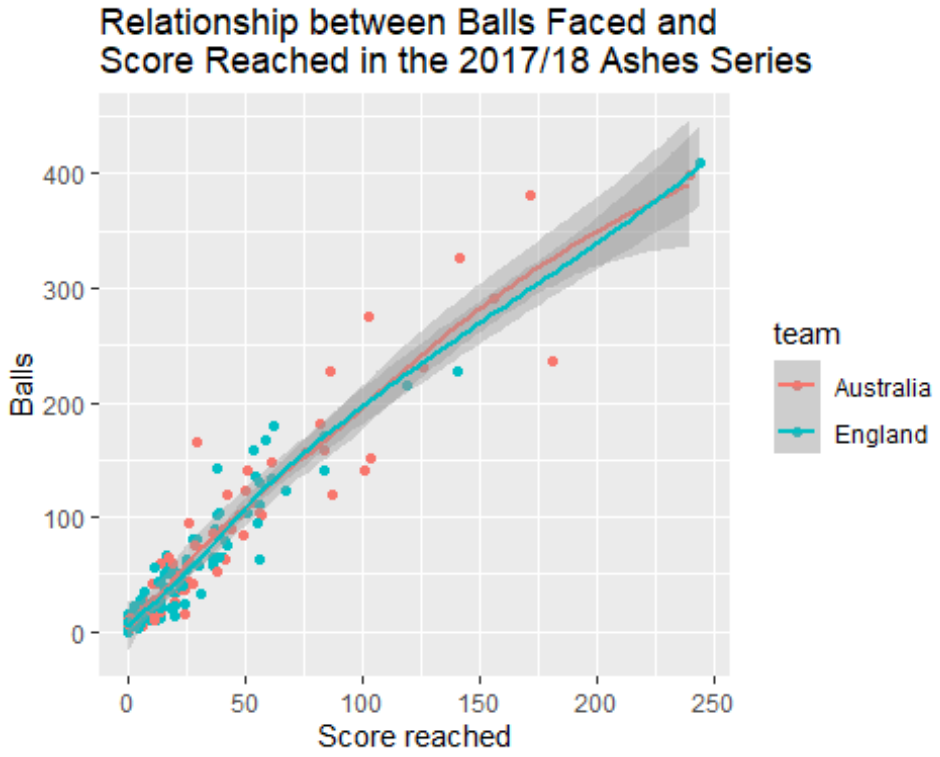

Figure 7: A scatterplot representing the relationship between scores reached and balls received in the 2017/18 Ashes series.

4.2

Describe the relationship between score and number of balls. Are players who face more balls likely to score more runs? [4 points]

There is a positive linear trend for both teams that indicates the more balls faced, the higher the score is likely to be; I've included a geom_smooth layer to clearly identify this trend (*figure 7*). There are a few things to consider:

1) The first is that you could refuse the first five balls and stay neutral provided you hit a six on the sixth ball. Its therefore quite easy to maintain the ratio of one ball to one run. In this way, there is a lot of room for batters to increase their scores and create a strong positive trendline.

2) A ball can only ever generate a neutral change in score of 0, it cannot reduce the offensive team's score; hence, the trendline can only ever be flat, positively trending, or non-existent in the event the team scores makes no runs. Statistically, the offensive team always has the advantage as every ball has the ability to increase the offensive teams score by 0, 1, 4, or 6. The only defense is to get the player out as soon as possible either by bowling them out or catching their hit.

3) This correlation pertains to this specific series. Where skill levels are approximately equivalent and there doesn't appear to be any contextual factors at first glance that drastically influenced players on game day. But consider that a great bowler could hit the stumps leaving the opposition team with a score of zero, or a defensive batter that could stay in without making a single run, four, or six. This would leave us with a very different correlation. The data from this specific series indicates that more balls will result in more runs, but it's important to be mindful that assumptions and context that apply here may not be true of other series.

4.3

Compute a new variable, scoring_rate, defined as the number of runs divided by the number of balls. Produce a scatterplot of scoring_rate against number of balls. [2 points]

```
scoring_rate_tibble <- ac %>%
  mutate(scoring_rates = runs_/balls_)
knitr::kable(head(scoring_rate_tibble), caption = "Table 7:Introduced a scori
ng rate column.")
```

Table 7: Introduced a scoring rate column.

innings	player	team	role	batting_order	runs_	balls_	scoring_rates
Test 1, Innings 1	Ali	England	all-rounder	6	38	102	0.3725490
Test 1, Innings 1	Anderson	England	bowler	11	5	9	0.5555556
Test 1, Innings 1	Bairstow	England	wicketkeeper	7	9	24	0.3750000
Test 1, Innings 1	Ball	England	bowler	10	14	11	1.2727273
Test 1, Innings 1	Bancroft	Australia	batsman	1	5	19	0.2631579
Test 1, Innings 1	Bird	Australia	bowler	NA	NA	NA	NA

```
ggplot(scoring_rate_tibble, aes( x = scoring_rates, y= balls_, col=team))+
  geom_point()+
  geom_smooth()+
  ggtitle("Relationship between Balls Faced and \nScoring Rate in the 2017/18
Ashes Series")+
  labs(x = "Scoring rates (Score/balls faced)", y="Balls faced")

nt).
```

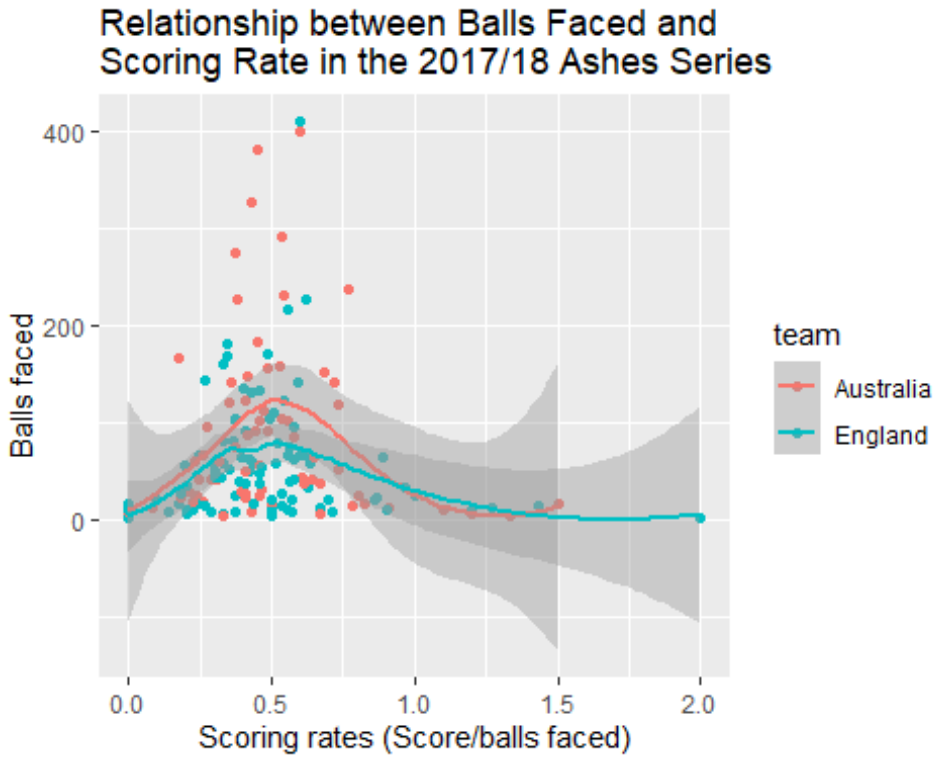

Figure 8: A scatterplot representing the relationship between the scoring rate and the balls received in the 2017/18 Ashes series

4.4

Is there a relationship between scoring rate and number of balls? Are players who face more balls likely to score runs more quickly? [2 points]

Scoring rate and balls faced do not appear to have a linear relationship. Logically, that makes perfect sense. Assuming a large number of balls and approximate skill-level equivalence, perhaps it would make sense for the first few balls to show an improvement in scoring rate as the player warmed up and gets their emotions in check. However, a linear trend would indicate that majority of players somehow improve or, in the case of a negative linear trend, get worse as they play. I wouldn't expect the best players Australia and England have to offer to do either of those things. Perhaps a new team over hundreds of games, but certainly not by the best of the best in a single test series.

Interestingly the geom_smooth function indicates that there is a negative quadratic relationship with a maximum at an approximate scoring rate of 0.5 at 100 balls. This shape indicates that scoring rates generally increase up until around the 100th ball. Indicating that the sooner the batter is out the better for the defending side; not very ground breaking. The cause is possibly to do with batting styles. Defensive players let more balls pass by, offensive players take more risks. Perhaps this just indicates the optimal batting style/ risk tolerance for timely ball to score conversion. In any case it's an interesting point for further investigation.

Question Five: Teams' roles

5.1

Produce a bar chart of the number of players on each team participating in the series, with segments coloured by the players' roles. [1 point]

```
ggplot(indiv_runs, aes(x=team, fill=role))+
  geom_bar()+
  ggtitle("Players per Team in the \n2017/218 Ashes Series")+
  labs(x = "Team", y= "Number of players")
```

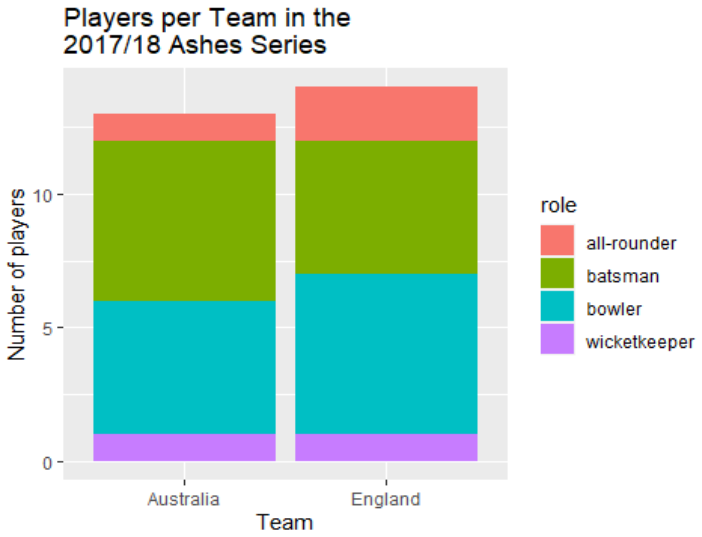

Figure 9: A bar chart representing the number of players on each team with colours indicating the proportion of player roles for that team in the 2017/18 Ashes series.

5.2

Produce a contingency table of the proportion of players from each team who play in each particular role. [2 points]

```
con_table <-  indiv_runs %>%
  group_by(role) %>%
  summarise(team, role, player)

#keeps 27 subjects and all variables required
con_table <- con_table %>%
  count(team, role)%>%
  spread(key  = "team", value = n)
#gives a table showing the total players in each roler per team
ct <- mutate(con_table, total = sum(Australia+England))
#adds a column for row totals
contingency_table <- ct%>%
  mutate(Aus=Australia/total, Eng= England/total)
#adds a column indicating the proportion of each
```

```
contingency_table <- contingency_table %>%
  mutate(Australia = NULL, England =NULL, total=NULL)
#removes unnecessary columns to reveal the...
knitr::kable(head(contingency_table), caption = "Table 8: Contingency table d
escribing the proportion of roles found in each team.")
```

Table 8: Contingency table describing the proportion of roles found in each team.

role	Aus	Eng
all-rounder	0.3333333	0.6666667
batsman	0.5454545	0.4545455
bowler	0.4545455	0.5454545
wicketkeeper	0.5000000	0.5000000

```
              #_____________Alternate method_____________#
install.packages("gmodels", repo = "https://cran.rstudio.com/bin/windows/Rtoo
ls/")

library(gmodels)
CrossTable(indiv_runs$role, indiv_runs$team)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  27
##
##
##                  | indiv_runs$team
## indiv_runs$role  | Australia |   England | Row Total |
## -----------------|-----------|-----------|-----------|
##      all-rounder |         1 |         2 |         3 |
##                  |     0.137 |     0.127 |           |
##                  |     0.333 |     0.667 |     0.111 |
##                  |     0.077 |     0.143 |           |
##                  |     0.037 |     0.074 |           |
## -----------------|-----------|-----------|-----------|
##          batsman |         6 |         5 |        11 |
##                  |     0.093 |     0.087 |           |
##                  |     0.545 |     0.455 |     0.407 |
##                  |     0.462 |     0.357 |           |
##                  |     0.222 |     0.185 |           |
## -----------------|-----------|-----------|-----------|
##           bowler |         5 |         6 |        11 |
##                  |     0.017 |     0.015 |           |
##                  |     0.455 |     0.545 |     0.407 |
##                  |     0.385 |     0.429 |           |
```

```
##                 |     0.185 |     0.222 |           |
## ----------------|-----------|-----------|-----------|
##    wicketkeeper |         1 |         1 |         2 |
##                 |     0.001 |     0.001 |           |
##                 |     0.500 |     0.500 |     0.074 |
##                 |     0.077 |     0.071 |           |
##                 |     0.037 |     0.037 |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |        13 |        14 |        27 |
##                 |     0.481 |     0.519 |           |
## ----------------|-----------|-----------|-----------|
##
##
```

```
#_____________________________________________________#
```

5.3

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. [2 points] [Total: 5 points]

The bar chart shows that Australia opted for an extra batsman, while England opted for an extra bowler. In doing so Australia had more batters. The English also had an additional all-rounder, thus having the highest proportion of them, and the larger team size. The contingency table puts numbers to that effect, indicating the proportion of player roles for each team (*table 8*). The proportion of batsman is higher for Australia, and bowler proportions are higher for England *(table 8*). Furthermore, the all-rounder row indicates the English doubled the number of all-rounders held by the Australians, and the number of keepers was equivalent (*table 8)*.

Question Six: Summary of Insights

Cricket Australia are interested in any insights you can bring with respect to the differences between the two teams, as well as any insights related to scoring. In plain English, write a summary of your key findings from Questions 2-5. Your response should be between 200-250 words. [3 points]

- Scoring rates probably don't provide a whole lot of meaning without observing how the score was accumulated (which balls were left, which were hit for four or six, which did the players run on). Knowing that will give you an indication of player batting styles, the influence of fatigue, and perhaps some insight into the optimal batting style. Furthermore, it will provide a frame work for defensive strategy in combatting the different types of playing styles.
- The optimal role proportions can't accurately be determined from just two teams alone, but with enough data on your team, other teams, and individual playing styles, you might be able to choose team role proportions that have a higher probability of countering the opposing team.
- No consideration has been made as yet to the position of the fielding team. This is perhaps the biggest variable in the game. Where are the defending team, how fast are they, how fast are the batsman, how fast can fielders throw the ball, and what kind of reach are they capable of? These questions could generate a heat map of locations that are likely to result in a batter going out if the ball is hit there. This could offer significant defensive potential and better offensive strategy. It would also enable the team to choose fielding locations that cover the field in a way that best defends against particular, and predictable, batting styles.

References:

R Core Team, 2021, *R: A language and environment for computing and statistical computing* R foundation for statistical computing Vienna, Austria.

Warnes, GR, Bolker, B, Lumley, T, Johnson, RC & SAIC-Frederick and Inc. 2018, 'gmodels: Various R Programming Tools for Model Fitting', <https://CRAN.R-project.org/package=gmodels>.

Wickham, H 2019, 'stringr: Simple, Consistent Wrappers for Common String Operations', R package version 1.4.0, <https://CRAN.R-project.org/package=stringr>.

Wickham, H, Averick, M, Bryan, J, Winston, C, McGowan, LDA, François, R, Grolemund, G, Hayes, A, Henry, L, Hester, J, Kuhn, M, Pedersen, TL, Miller, E, Bache, SM, Müller, K, Ooms, J, Robinson, D, Seidel, DP, Spinu, V, Takahashi, K, Vaughan, D, Wilke, C, Woo, K & Yutani, H 2019, 'Welcome to the {tidyverse}', *Journal of Open Source Software*, vol. 4, no. 43, p. 1686.

Wickham, H, François, R, Henry, L & Müller, K 2021, 'dplyr: A Grammar of Data Manipulation', R package version 1.0.7, <https://CRAN.R-project.org/package=dplyr>.

```
 citation()

citation("tidyverse")

citation("dplyr")

citation("stringr")

citation("gmodels")
```