Assignment1

Chia-Hao Lo

September 28th 2024

Setup

```
#Load the required packages
library(tidyverse)
library(inspectdf)
library(caret)
library(moments)
library(tidymodels)
library(modelr)
library(ISLR)
library(car)
```

Q1. Loading the data

We need to calculate our ysn to satisfy the delivery specification

```
# Your student number goes here
ysn = 1907385
# Calculate your student number modulo 3
mod4 <- ysn %% 4
mod4
```

```
## [1] 1
```

[After calculating ysn we get 1. Due to deliverable specification that ysn = 1 mod 4 use the Indian data, we will use Indian data as our dataframe]

```
filename <- paste0("a1_1907385_3", ".pdf")
filename
```

```
## [1] "a1_1907385_3.pdf"
```

```
# Read in the data using the correct tidyverse command

ind1 <- read_csv("./data/India.csv")
ind1 <- as_tibble(ind1)

# Display the first 10 lines of the data
head(ind1, 10)
```

```
## # A tibble: 10 x 8
##    Player          RightHanded `2000`          `2001` `2002` `2003` `2004` `2005`
##    <chr>           <chr>       <chr>           <chr>  <chr>  <chr>  <chr>  <chr>
##  1 A Chopra        yes         0 innings, 0 ~ 0 inn~ 0 inn~ 10 in~ 9 inn~ 0 inn~
```

```
##  2 A Jadeja        yes         2 innings, 2 ~ 0 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
##  3 A Kumble        yes         6 innings, 5 ~ 6 inn~ 12 in~ 4 inn~ 14 in~ 11 in~
##  4 A Nehra         yes         0 innings, 0 ~ 5 inn~ 15 in~ 4 inn~ 1 inn~ 0 inn~
##  5 A Ratra         yes         0 innings, 0 ~ 0 inn~ 10 in~ 0 inn~ 0 inn~ 0 inn~
##  6 AB Agarkar      yes         6 innings, 5 ~ 5 inn~ 8 inn~ 5 inn~ 5 inn~ 4 inn~
##  7 BKV Prasad      yes         2 innings, 1 ~ 6 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
##  8 D Dasgupta      yes         0 innings, 0 ~ 9 inn~ 4 inn~ 0 inn~ 0 inn~ 0 inn~
##  9 G Gambhir       no          0 innings, 0 ~ 0 inn~ 0 inn~ 0 inn~ 7 inn~ 14 in~
## 10 Harbhajan Singh yes         0 innings, 0 ~ 18 in~ 20 in~ 2 inn~ 9 inn~ 10 in~
```

Q2. The dimensions of the data set

For this part, we will use dim() function to get a list of the form.

```
#Dimension
dim(ind1)
```

```
## [1] 50  8
```

The dimensions of the data frame is [50 rows, 8 columns]

Q3. Random permutation of the rows

From the delivery specification, we know seed equals to ysn, we gt ysn from Q1 and ysn = 1. Then we permute all rows into data set.

```
#
set.seed(ysn)
ind1_permuted <- sample_n(ind1, 50)
head(ind1_permuted, 10)
```

```
## # A tibble: 10 x 8
##    Player            RightHanded `2000`       `2001` `2002` `2003` `2004` `2005`
##    <chr>             <chr>       <chr>        <chr>  <chr>  <chr>  <chr>  <chr>
##  1 R Vijay Bharadwaj yes         1 innings, ~ 0 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
##  2 J Srinath         yes         7 innings, ~ 11 in~ 12 in~ 0 inn~ 0 inn~ 0 inn~
##  3 NM Kulkarni       no          0 innings, ~ 1 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
##  4 Sarandeep Singh   yes         0 innings, ~ 1 inn~ 1 inn~ 0 inn~ 0 inn~ 0 inn~
##  5 SC Ganguly        no          10 innings,~ 23 in~ 25 in~ 7 inn~ 9 inn~ 10 in~
##  6 N Chopra          yes         2 innings, ~ 0 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
##  7 BKV Prasad        yes         2 innings, ~ 6 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
##  8 test1test1        x           test         test   test   test   test   test
##  9 Harbhajan Singh   yes         0 innings, ~ 18 in~ 20 in~ 2 inn~ 9 inn~ 10 in~
## 10 M Kartik          no          5 innings, ~ 0 inn~ 0 inn~ 0 inn~ 5 inn~ 0 inn~
```

The result shows the first 10 rows.

Q4. Adding an extra column of row numbers

We add new rows and use relocate() to set this row to the first row.

```
#
ind1_permuted1 <- ind1_permuted %>%
  mutate(Rows = row_number()) %>% relocate("Rows", .before = Player)

head(ind1_permuted1, 10)
```

```
## # A tibble: 10 x 9
##     Rows Player            RightHanded `2000` `2001` `2002` `2003` `2004` `2005`
##    <int> <chr>             <chr>       <chr>  <chr>  <chr>  <chr>  <chr>  <chr>
##  1     1 R Vijay Bharadwaj yes         1 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
##  2     2 J Srinath         yes         7 inn~ 11 in~ 12 in~ 0 inn~ 0 inn~ 0 inn~
##  3     3 NM Kulkarni       no          0 inn~ 1 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
##  4     4 Sarandeep Singh   yes         0 inn~ 1 inn~ 1 inn~ 0 inn~ 0 inn~ 0 inn~
##  5     5 SC Ganguly        no          10 in~ 23 in~ 25 in~ 7 inn~ 9 inn~ 10 in~
##  6     6 N Chopra          yes         2 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
##  7     7 BKV Prasad        yes         2 inn~ 6 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
##  8     8 test1test1        x           test   test   test   test   test   test
##  9     9 Harbhajan Singh   yes         0 inn~ 18 in~ 20 in~ 2 inn~ 9 inn~ 10 in~
## 10    10 M Kartik          no          5 inn~ 0 inn~ 0 inn~ 0 inn~ 5 inn~ 0 inn~
```

Q5 Data cleaning

Here we test if the data set contains the duplicated data.

```
# Check if the data has some duplicated rows
ind1_clean <- ind1_permuted1 %>%
  distinct()
sum(duplicated(ind1_clean))
```

```
## [1] 0
```

Here we delete test data

```
ind1_clean1 <- ind1_clean %>%
  filter(!(Player == "test1test1")
         )
```

Here we convert negative data to positive data.

```
#To get absolute value for whole data
ind1_clean1[ind1_clean1$Player == "IK Pathan", "2005"] <-
  str_replace(ind1_clean1[ind1_clean1$Player == "IK Pathan", "2005"], "-", "")
ind1_clean1
```

```
## # A tibble: 49 x 9
##     Rows Player            RightHanded `2000` `2001` `2002` `2003` `2004` `2005`
##    <int> <chr>             <chr>       <chr>  <chr>  <chr>  <chr>  <chr>  <chr>
##  1     1 R Vijay Bharadwaj yes         1 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
##  2     2 J Srinath         yes         7 inn~ 11 in~ 12 in~ 0 inn~ 0 inn~ 0 inn~
##  3     3 NM Kulkarni       no          0 inn~ 1 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
##  4     4 Sarandeep Singh   yes         0 inn~ 1 inn~ 1 inn~ 0 inn~ 0 inn~ 0 inn~
##  5     5 SC Ganguly        no          10 in~ 23 in~ 25 in~ 7 inn~ 9 inn~ 10 in~
##  6     6 N Chopra          yes         2 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
##  7     7 BKV Prasad        yes         2 inn~ 6 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
##  8     9 Harbhajan Singh   yes         0 inn~ 18 in~ 20 in~ 2 inn~ 9 inn~ 10 in~
##  9    10 M Kartik          no          5 inn~ 0 inn~ 0 inn~ 0 inn~ 5 inn~ 0 inn~
## 10    11 HH Kanitkar       no          2 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
## # i 39 more rows
```

Here we detect if here is any NA data.

```
#To check if there is any missing values in any columns
inspect_na(ind1_clean1)
```

```
## # A tibble: 9 x 3
##   col_name      cnt  pcnt
##   <chr>       <int> <dbl>
## 1 Rows            0     0
## 2 Player          0     0
## 3 RightHanded     0     0
## 4 2000            0     0
## 5 2001            0     0
## 6 2002            0     0
## 7 2003            0     0
## 8 2004            0     0
## 9 2005            0     0
```

Here we need to convert word zero to number 0

```
ind1_clean2 <- ind1_clean1 %>%
  mutate(`2002` = str_replace(`2002`, "zero", "0"))
print(ind1_clean2[ind1_clean2$Player == "T Yohannan", "2002"])
```

```
## # A tibble: 1 x 1
##   `2002`
##   <chr>
## 1 2 innings, 0 outs, 8 runs, 9 balls
```

Delete player NM Kulkarni row because in 2001 data is impossible for one innings to get 9999 runs.

```
ind1_clean3 <- ind1_clean2 %>%
  filter(!(Player == "NM Kulkarni")
  )

print(ind1_clean3 %>% filter(Player == "NM Kulkarni"))
```

```
## # A tibble: 0 x 9
## # i 9 variables: Rows <int>, Player <chr>, RightHanded <chr>, 2000 <chr>,
## #   2001 <chr>, 2002 <chr>, 2003 <chr>, 2004 <chr>, 2005 <chr>
```

To find the values in RightHanded column that is not yes or no

```
invalid_values <- ind1_clean3[!(ind1_clean3$RightHanded %in% c("yes", "no")),]
invalid_values
```

```
## # A tibble: 2 x 9
##    Rows Player   RightHanded `2000`           `2001` `2002` `2003` `2004` `2005`
##   <int> <chr>    <chr>       <chr>            <chr>  <chr>  <chr>  <chr>  <chr>
## 1    31 L Balaji Y           0 innings, 0 ou~ 0 inn~ 0 inn~ 1 inn~ 3 inn~ 5 inn~
## 2    39 SS Dighe Y           0 innings, 0 ou~ 10 in~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
```

Here we found out there are two rows need to fix. First we change rows in RightHanded column contain Y to yes, like other data.

```
#Change Y to yes
ind1_clean3$RightHanded[ind1_clean3$RightHanded == "Y"] <- "yes"
ind1_clean3
```

```
## # A tibble: 48 x 9
##     Rows Player            RightHanded `2000` `2001` `2002` `2003` `2004` `2005`
```

```
##    <int> <chr>             <chr>       <chr>  <chr>  <chr>  <chr>  <chr>  <chr>
##  1     1 R Vijay Bharadwaj yes         1 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
##  2     2 J Srinath         yes         7 inn~ 11 in~ 12 in~ 0 inn~ 0 inn~ 0 inn~
##  3     4 Sarandeep Singh   yes         0 inn~ 1 inn~ 1 inn~ 0 inn~ 0 inn~ 0 inn~
##  4     5 SC Ganguly        no          10 in~ 23 in~ 25 in~ 7 inn~ 9 inn~ 10 in~
##  5     6 N Chopra          yes         2 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
##  6     7 BKV Prasad        yes         2 inn~ 6 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
##  7     9 Harbhajan Singh   yes         0 inn~ 18 in~ 20 in~ 2 inn~ 9 inn~ 10 in~
##  8    10 M Kartik          no          5 inn~ 0 inn~ 0 inn~ 0 inn~ 5 inn~ 0 inn~
##  9    11 HH Kanitkar       no          2 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~ 0 inn~
## 10    12 T Yohannan        yes         0 inn~ 2 inn~ 2 inn~ 0 inn~ 0 inn~ 0 inn~
## # i 38 more rows
```

After we look at this row we know this row is test row and there is no valued data. So we will remove it.

After cleaning the database, we start to tidy the data. First, we convert 2000-2005 columns to two new columns year and performance.

Q6 Tidy data

(a)

```
ind1_convert <- ind1_clean3 %>%
  gather(key = "year", value = "performance", starts_with("200"))

ind1_convert
```

```
## # A tibble: 288 x 5
##     Rows Player            RightHanded year  performance
##    <int> <chr>             <chr>       <chr> <chr>
##  1     1 R Vijay Bharadwaj yes         2000  1 innings, 1 outs, 6 runs, 35 balls
##  2     2 J Srinath         yes         2000  7 innings, 5 outs, 25 runs, 80 bal~
##  3     4 Sarandeep Singh   yes         2000  0 innings, 0 outs, 0 runs, 0 balls
##  4     5 SC Ganguly        no          2000  10 innings, 9 outs, 279 runs, 540 ~
##  5     6 N Chopra          yes         2000  2 innings, 2 outs, 7 runs, 70 balls
##  6     7 BKV Prasad        yes         2000  2 innings, 1 outs, 4 runs, 9 balls
##  7     9 Harbhajan Singh   yes         2000  0 innings, 0 outs, 0 runs, 0 balls
##  8    10 M Kartik          no          2000  5 innings, 5 outs, 61 runs, 183 ba~
##  9    11 HH Kanitkar       no          2000  2 innings, 2 outs, 18 runs, 86 bal~
## 10    12 T Yohannan        yes         2000  0 innings, 0 outs, 0 runs, 0 balls
## # i 278 more rows
```

Here we separate the performance to four new columns.

(b)

```
details <- str_match(ind1_convert$performance, "(\\d+) innings, (\\d+) outs, (\\d+) runs, (\\d+) balls"
ind1_performance <- ind1_convert %>%
  mutate(
    Innings = as.numeric(details[,2]),
    Outs = as.numeric(details[,3]),
    Runs = as.numeric(details[,4]),
    Balls = as.numeric(details[,5])
```

```
  )
ind1_performance
```

```
## # A tibble: 288 x 9
##     Rows Player          RightHanded year  performance Innings  Outs  Runs Balls
##    <int> <chr>           <chr>       <chr> <chr>         <dbl> <dbl> <dbl> <dbl>
##  1     1 R Vijay Bharad~ yes         2000  1 innings,~       1     1     6    35
##  2     2 J Srinath       yes         2000  7 innings,~       7     5    25    80
##  3     4 Sarandeep Singh yes         2000  0 innings,~       0     0     0     0
##  4     5 SC Ganguly      no          2000  10 innings~      10     9   279   540
##  5     6 N Chopra        yes         2000  2 innings,~       2     2     7    70
##  6     7 BKV Prasad      yes         2000  2 innings,~       2     1     4     9
##  7     9 Harbhajan Singh yes         2000  0 innings,~       0     0     0     0
##  8    10 M Kartik        no          2000  5 innings,~       5     5    61   183
##  9    11 HH Kanitkar     no          2000  2 innings,~       2     2    18    86
## 10    12 T Yohannan      yes         2000  0 innings,~       0     0     0     0
## # i 278 more rows
```

After we get four new columns then we can delete performance column.

(c)

```
ind1_cleaned <- ind1_performance %>%
  select(-performance)
ind1_cleaned
```

```
## # A tibble: 288 x 8
##     Rows Player            RightHanded year  Innings  Outs  Runs Balls
##    <int> <chr>             <chr>       <chr>   <dbl> <dbl> <dbl> <dbl>
##  1     1 R Vijay Bharadwaj yes         2000        1     1     6    35
##  2     2 J Srinath         yes         2000        7     5    25    80
##  3     4 Sarandeep Singh   yes         2000        0     0     0     0
##  4     5 SC Ganguly        no          2000       10     9   279   540
##  5     6 N Chopra          yes         2000        2     2     7    70
##  6     7 BKV Prasad        yes         2000        2     1     4     9
##  7     9 Harbhajan Singh   yes         2000        0     0     0     0
##  8    10 M Kartik          no          2000        5     5    61   183
##  9    11 HH Kanitkar       no          2000        2     2    18    86
## 10    12 T Yohannan        yes         2000        0     0     0     0
## # i 278 more rows
```

Here we add T_rows with the row numbers

(d)

```
ind1_cleaned1 <- ind1_cleaned %>%
  mutate(T_rows = row_number()) %>%
  relocate(T_rows, .before = Player)
head(ind1_cleaned1, 10)
```

```
## # A tibble: 10 x 9
##     Rows T_rows Player            RightHanded year  Innings  Outs  Runs Balls
##    <int>  <int> <chr>             <chr>       <chr>   <dbl> <dbl> <dbl> <dbl>
```

```
##  1     1      1 R Vijay Bharadwaj yes          2000        1     1     6    35
##  2     2      2 J Srinath         yes          2000        7     5    25    80
##  3     4      3 Sarandeep Singh   yes          2000        0     0     0     0
##  4     5      4 SC Ganguly        no           2000       10     9   279   540
##  5     6      5 N Chopra          yes          2000        2     2     7    70
##  6     7      6 BKV Prasad        yes          2000        2     1     4     9
##  7     9      7 Harbhajan Singh   yes          2000        0     0     0     0
##  8    10      8 M Kartik          no           2000        5     5    61   183
##  9    11      9 HH Kanitkar       no           2000        2     2    18    86
## 10    12     10 T Yohannan        yes          2000        0     0     0     0
```

Q7 Data types

.Rows: Categorical Ordinal(each row number follows the order and it is for category the order of the players)

.T_rows: Categorical Ordinal(each row number follows the order and it is for category the order of the rows)

.Player: Categorical Nominal(names of player, which are labels also no natural order for player names)

.RightHanded: Categorical Nominal(it shows if the player is righthanded, this is a categorical nominal variable, which is yes or no)

.year: Categorical Ordinal(represents a sequence of time)

.Innings:Quantitative Discrete(the counts of innings played in a given year)

.Outs:Quantitative Discrete(counts how many times a player was out)

.Runs:Quantitative Discrete(total scores of runs scored by a player)

.Balls:Quantitative Discrete(certain counts of a player faced the total number of balls)

Q8 Tame data

Here we will follow Tame data from my uni material steps, first we lower case, then we convert the data type to question 7 data type we found.

```r
ind1_taming <- ind1_cleaned1 %>%
  rename_with(tolower)

ind1_taming$rows <- as.ordered(ind1_taming$rows)

ind1_taming$t_rows <- as.ordered(ind1_taming$t_rows)

ind1_taming$year <- as.ordered(ind1_taming$year)

ind1_taming$player <- as.factor(ind1_taming$player)
head(ind1_taming, 10)
```

```
## # A tibble: 10 x 9
##    rows  t_rows player            righthanded year  innings  outs  runs balls
##    <ord> <ord>  <fct>             <chr>       <ord>   <dbl> <dbl> <dbl> <dbl>
##  1 1     1      R Vijay Bharadwaj yes         2000        1     1     6    35
##  2 2     2      J Srinath         yes         2000        7     5    25    80
##  3 4     3      Sarandeep Singh   yes         2000        0     0     0     0
##  4 5     4      SC Ganguly        no          2000       10     9   279   540
##  5 6     5      N Chopra          yes         2000        2     2     7    70
##  6 7     6      BKV Prasad        yes         2000        2     1     4     9
```

```
##  7 9     7      Harbhajan Singh   yes          2000        0     0     0     0
##  8 10    8      M Kartik          no           2000        5     5    61   183
##  9 11    9      HH Kanitkar       no           2000        2     2    18    86
## 10 12    10     T Yohannan        yes          2000        0     0     0     0
```

```
dim(ind1_taming)
```

```
## [1] 288   9
```

Here we filter the data is non-zero, we remove the balls are equal to zero

Q9 Non-zero ball data

```
ind1_balls <- ind1_taming %>%
  filter(balls > 0)
head(ind1_balls, 10)
```

```
## # A tibble: 10 x 9
##    rows  t_rows player            righthanded year  innings  outs  runs balls
##    <ord> <ord>  <fct>             <chr>       <ord>   <dbl> <dbl> <dbl> <dbl>
##  1 1     1      R Vijay Bharadwaj yes         2000        1     1     6    35
##  2 2     2      J Srinath         yes         2000        7     5    25    80
##  3 5     4      SC Ganguly        no          2000       10     9   279   540
##  4 6     5      N Chopra          yes         2000        2     2     7    70
##  5 7     6      BKV Prasad        yes         2000        2     1     4     9
##  6 10    8      M Kartik          no          2000        5     5    61   183
##  7 11    9      HH Kanitkar       no          2000        2     2    18    86
##  8 14    12     AB Agarkar        yes         2000        6     5    90   178
##  9 16    14     SR Tendulkar      yes         2000       10     9   575   954
## 10 17    15     SB Joshi          no          2000        2     2   119   198
```

```
dim(ind1_balls)
```

```
## [1] 121   9
```

Here we set the correct seed again and set 70 rows

Q10 Random subset

(a)

```
set.seed(ysn)
ind1_sample <- sample_n(ind1_balls, 70)
```

Here we sort the rows by tidy row numbers # (b)

```
ind1_sorted <- ind1_balls %>%
  arrange(t_rows)
head(ind1_sorted, 10)
```

```
## # A tibble: 10 x 9
##    rows  t_rows player            righthanded year  innings  outs  runs balls
##    <ord> <ord>  <fct>             <chr>       <ord>   <dbl> <dbl> <dbl> <dbl>
##  1 1     1      R Vijay Bharadwaj yes         2000        1     1     6    35
##  2 2     2      J Srinath         yes         2000        7     5    25    80
```

```
##  3 5     4      SC Ganguly        no          2000       10     9   279   540
##  4 6     5      N Chopra          yes         2000        2     2     7    70
##  5 7     6      BKV Prasad        yes         2000        2     1     4     9
##  6 10    8      M Kartik          no          2000        5     5    61   183
##  7 11    9      HH Kanitkar       no          2000        2     2    18    86
##  8 14    12     AB Agarkar        yes         2000        6     5    90   178
##  9 16    14     SR Tendulkar      yes         2000       10     9   575   954
## 10 17    15     SB Joshi          no          2000        2     2   119   198
```

```
dim(ind1_sorted)
```

```
## [1] 121   9
```

Here we calculate the percentage of innings and he number of runs scored per ball, then change two columns' location

Q11 pct_out & run_rate

(a)

```
ind1_new_columns <- ind1_sorted %>%
  mutate(
    pct_out = round((outs/innings)*100, 2),
    run_rate = round(runs/balls, 3)) %>%
  relocate(pct_out, run_rate, .after = year)
#head(ind1_new_columns, 10)
ind1_new_columns
```

```
## # A tibble: 121 x 11
##    rows  t_rows player    righthanded year  pct_out run_rate innings  outs  runs
##    <ord> <ord>  <fct>     <chr>       <ord>   <dbl>    <dbl>   <dbl> <dbl> <dbl>
##  1 1     1      R Vijay ~ yes         2000    100      0.171       1     1     6
##  2 2     2      J Srinath yes         2000     71.4    0.312       7     5    25
##  3 5     4      SC Gangu~ no          2000     90      0.517      10     9   279
##  4 6     5      N Chopra  yes         2000    100      0.1         2     2     7
##  5 7     6      BKV Pras~ yes         2000     50      0.444       2     1     4
##  6 10    8      M Kartik  no          2000    100      0.333       5     5    61
##  7 11    9      HH Kanit~ no          2000    100      0.209       2     2    18
##  8 14    12     AB Agark~ yes         2000     83.3    0.506       6     5    90
##  9 16    14     SR Tendu~ yes         2000     90      0.603      10     9   575
## 10 17    15     SB Joshi  no          2000    100      0.601       2     2   119
## # i 111 more rows
## # i 1 more variable: balls <dbl>
```

.pct_out: Quantitative Continuous(this is a numeric and presents by percentage)

.run_rate: Quantitative Continuous(this represents the runs score per ball and it can be any numeric value)

```
dim(ind1_new_columns)
```

```
## [1] 121  11
```

(b)

For the percentage of outs, it is calculate outs/innings, it is the same as pct_out. Thus we use min() function to find the lowest percentage of outs.

```
min_pct_out = ind1_new_columns %>%
  filter(pct_out == min(pct_out)) %>%
  select(player, year, pct_out)
min_pct_out
```

```
## # A tibble: 6 x 3
##   player          year  pct_out
##   <fct>           <ord>   <dbl>
## 1 Z Khan          2000        0
## 2 V Dahiya        2000        0
## 3 T Yohannan      2001        0
## 4 Sarandeep Singh 2002        0
## 5 T Yohannan      2002        0
## 6 A Nehra         2004        0
```

From the results, there are 6 players with the lowest percentage of outs, which are same value of 0 %.

We need to find the highest run rate, which is run_rate's dataset. We use max() function to find the highest value.

```
max_run_rate <- ind1_new_columns %>%
  filter(run_rate == max(run_rate)) %>%
  select(player, year, run_rate)
max_run_rate
```

```
## # A tibble: 1 x 3
##   player      year  run_rate
##   <fct>       <ord>    <dbl>
## 1 IR Siddiqui 2001     0.967
```

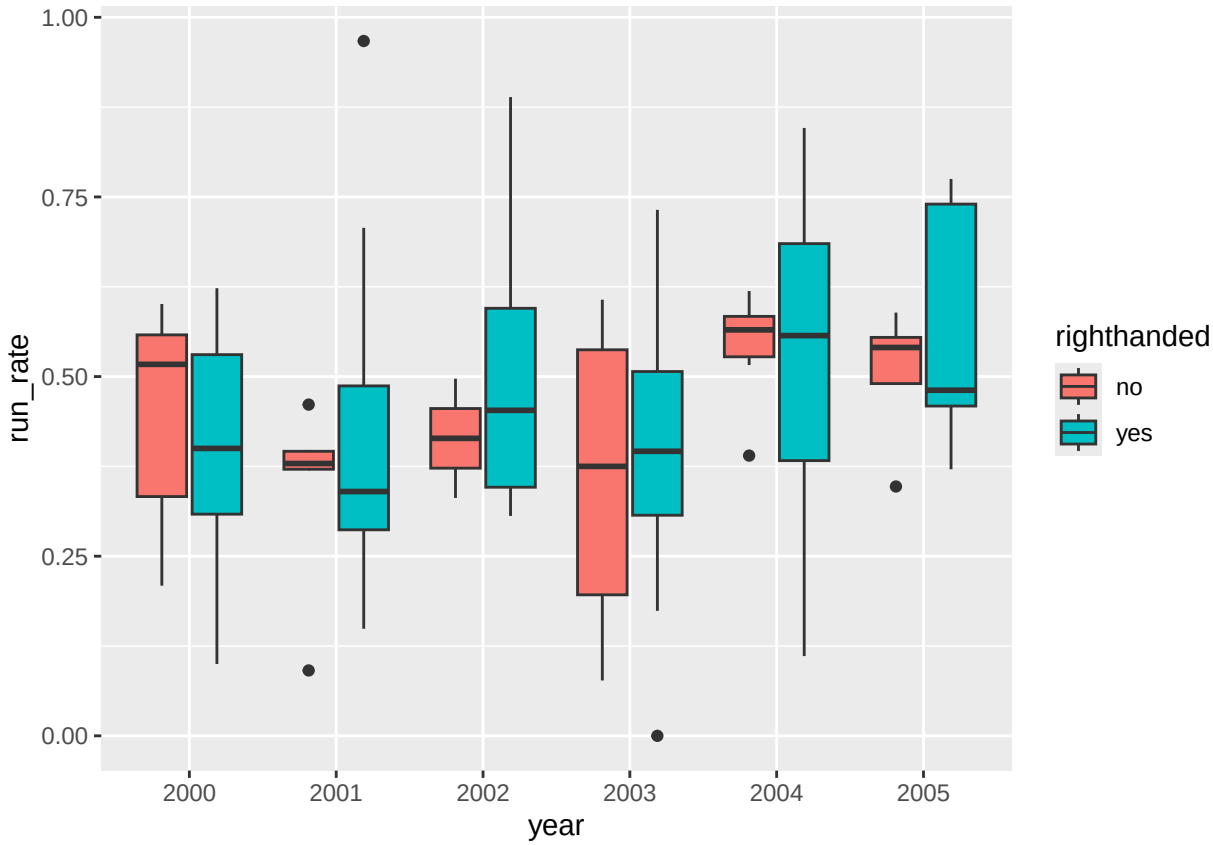
For the highest run rate, the result shows in 2001, and the value is 0.967.

Here we change the year data type and perform the boxplot

Q12 Boxplot using year

```
ind1_new_columns1 <- mutate(ind1_new_columns, year = factor(year, ordered = TRUE))

ggplot(ind1_new_columns1, aes(x = year, y = run_rate, fill = righthanded)) +
  geom_boxplot()
```

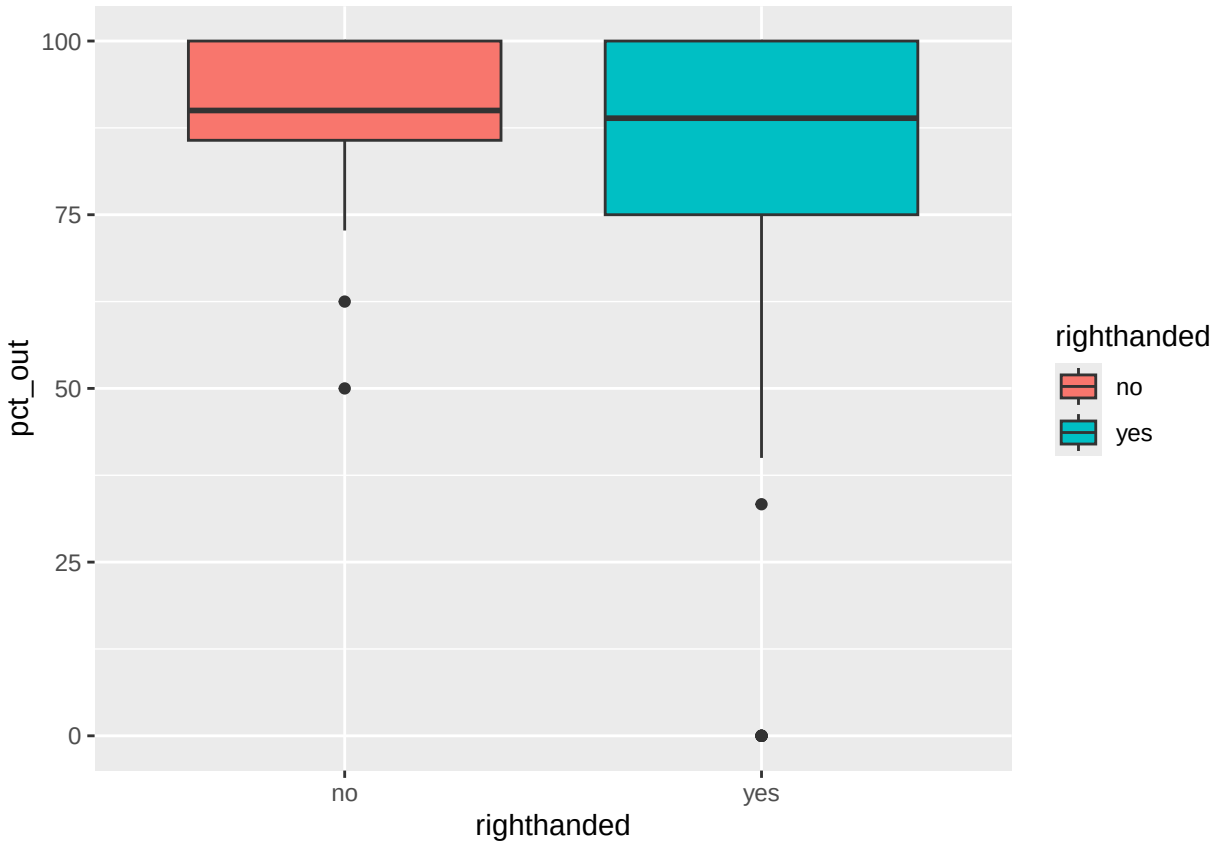


Here we also perform two boxplots

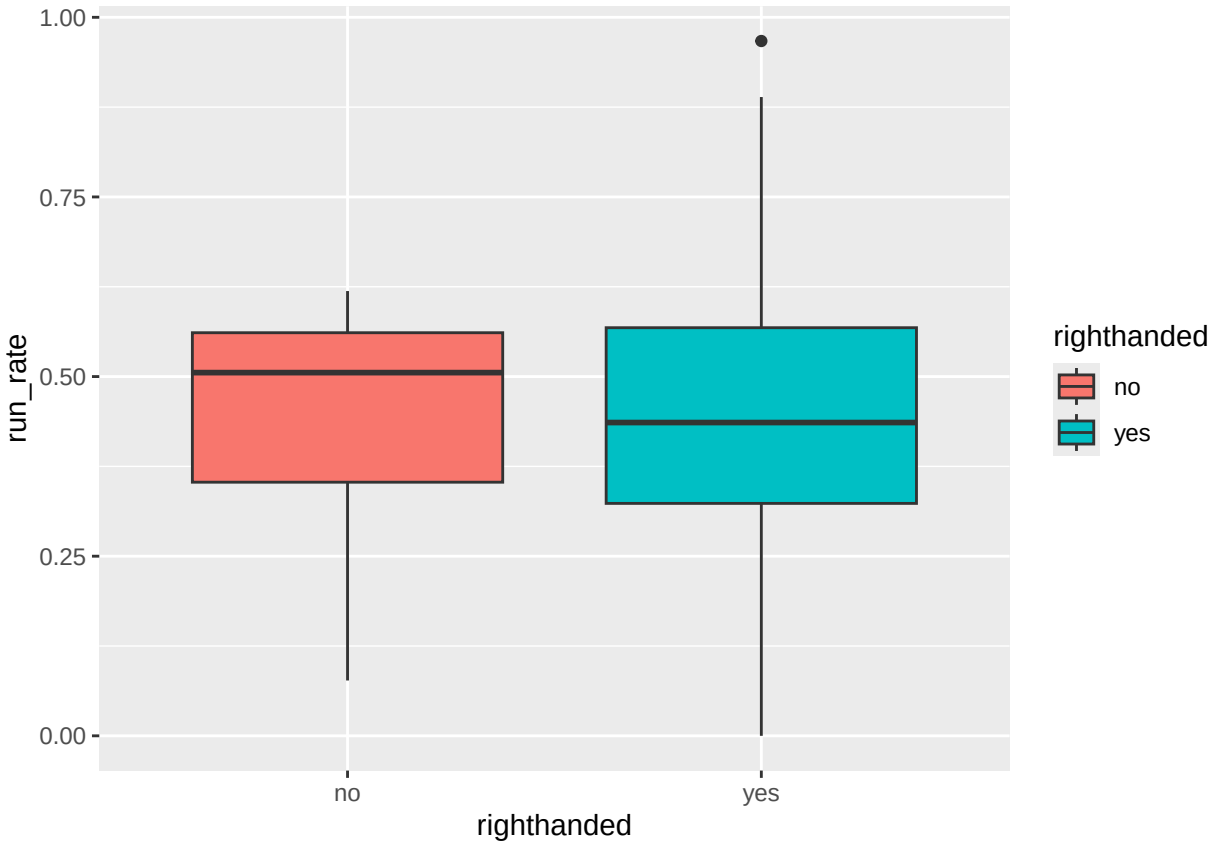
Q13 Side-by-side boxplots

```
boxplot_pct_out <- ggplot(ind1_new_columns1, aes(x = righthanded, y = pct_out, fill = righthanded)) +
  geom_boxplot()

boxplot_pct_out
```




```
boxplot_run_rate <- ggplot(ind1_new_columns1, aes(x = righthanded, y = run_rate, fill = righthanded)) +
  geom_boxplot()
boxplot_run_rate
```



Q14 Analysis

Based on the analysis in Q11, we found that players with a lower percentage of outs (pct_out) generally perform blanded. For example, T Yohannan had the lowest percentage of outs at 0%, his data mostly is 0 and it means he has lower priority playing rate. In contrast, players with a higher percentage of outs were more likely to have higher run rates, as their innings were higher.

In the box plots, we can see the patterns based on handedness. right-handed players had a wider variation in run_rate across years compared to left-handed players. For example, in the year 2004, right-handed players had a significantly higher spread in run_rate. For left-handed players showed more consistency, like 2002, the run_rate distribution was more compact.

Additionally, the box plots of pct_out and run_rate by handedness shows right-handed players seemed to perform better on average, as shown by a higher median run_rate and a generally lower pct_out. This suggests that right-handed players may have an advantage or perhaps more opportunities in the dataset used.