MATHS 7107 Data Taming Tutorial Week 8

Chang Dong

2023-03-20

1. Load the libraries.

```
pacman::p_load("tidymodels","tidyverse","stringr")
```

2. Read the data.

```
origin_runs <-read.csv("test_runs_original.csv")
runs <- read.csv("test_runs.csv")
```

3. Extract the country code from the player variable and create a new variable called Country. hint: You can use "\((.+?)\)" regular expression.

```
origin_runs <-origin_runs %>% mutate(Country = str_match(origin_runs$Player,pattern = "\\((.+?)\\)")[,2
origin_runs %>% head()
```

```
##   X               Player      Span Mat Inns NO  Runs   HS   Ave X100 X50 X0
## 1 1 SR Tendulkar (INDIA) 1989-2013 200  329 33 15921 248* 53.78   51  68 14
## 2 2     RT Ponting (AUS) 1995-2012 168  287 29 13378  257 51.85   41  62 17
## 3 3   JH Kallis (ICC/SA) 1995-2013 166  280 40 13289  224 55.37   45  58 16
## 4 4 R Dravid (ICC/INDIA) 1996-2012 164  286 32 13288  270 52.31   36  63  8
## 5 5        AN Cook (ENG) 2006-2018 161  291 16 12472  294 45.35   33  57  9
## 6 6   KC Sangakkara (SL) 2000-2015 134  233 17 12400  319 57.40   38  52 11
##     Country
## 1     INDIA
## 2       AUS
## 3    ICC/SA
## 4 ICC/INDIA
## 5       ENG
## 6        SL
```

4.Remove the country code from the players name.

```
origin_runs$Player = str_replace(origin_runs$Player,pattern = "\\s*\\((.+?)\\)", "")
origin_runs %>% head()
```

```
##   X        Player      Span Mat Inns NO  Runs   HS   Ave X100 X50 X0   Country
## 1 1  SR Tendulkar 1989-2013 200  329 33 15921 248* 53.78   51  68 14     INDIA
## 2 2    RT Ponting 1995-2012 168  287 29 13378  257 51.85   41  62 17       AUS
## 3 3     JH Kallis 1995-2013 166  280 40 13289  224 55.37   45  58 16    ICC/SA
## 4 4      R Dravid 1996-2012 164  286 32 13288  270 52.31   36  63  8 ICC/INDIA
## 5 5       AN Cook 2006-2018 161  291 16 12472  294 45.35   33  57  9       ENG
```

```
## 6 6 KC Sangakkara 2000-2015 134  233 17 12400  319 57.40   38  52 11        SL
```

5.Extract only the number from the HS (Highest score) variable?

```
origin_runs$HS = str_remove(origin_runs$HS,pattern = "\\*")
origin_runs %>% head()
```

```
##   X        Player      Span Mat Inns NO  Runs  HS   Ave X100 X50 X0   Country
## 1 1  SR Tendulkar 1989-2013 200  329 33 15921 248 53.78   51  68 14     INDIA
## 2 2    RT Ponting 1995-2012 168  287 29 13378 257 51.85   41  62 17       AUS
## 3 3     JH Kallis 1995-2013 166  280 40 13289 224 55.37   45  58 16    ICC/SA
## 4 4      R Dravid 1996-2012 164  286 32 13288 270 52.31   36  63  8 ICC/INDIA
## 5 5       AN Cook 2006-2018 161  291 16 12472 294 45.35   33  57  9       ENG
## 6 6 KC Sangakkara 2000-2015 134  233 17 12400 319 57.40   38  52 11        SL
```

6. Change variable names as follow

100 as Centuries 50 as Fifties 0 as Zeros

```
origin_runs<- origin_runs %>% rename(Centuries =  X100,
                                     Fifties = X50,
                                     Zeros = X0)
origin_runs %>% head()
```

```
##   X        Player      Span Mat Inns NO  Runs  HS   Ave Centuries Fifties Zeros
## 1 1  SR Tendulkar 1989-2013 200  329 33 15921 248 53.78        51      68    14
## 2 2    RT Ponting 1995-2012 168  287 29 13378 257 51.85        41      62    17
## 3 3     JH Kallis 1995-2013 166  280 40 13289 224 55.37        45      58    16
## 4 4      R Dravid 1996-2012 164  286 32 13288 270 52.31        36      63     8
## 5 5       AN Cook 2006-2018 161  291 16 12472 294 45.35        33      57     9
## 6 6 KC Sangakkara 2000-2015 134  233 17 12400 319 57.40        38      52    11
##     Country
## 1     INDIA
## 2       AUS
## 3    ICC/SA
## 4 ICC/INDIA
## 5       ENG
## 6        SL
```

7. Recode the relevant variables as factors and integers

```
origin_runs$HS <- as.numeric(origin_runs$HS)
origin_runs$Country <- as.factor((origin_runs$Country))
```

8. Recode the factors using fct_recode() such that there are no typographical errors in the Country variable.

```
levels(origin_runs$Country) <- fct_recode(levels(origin_runs$Country),
"INDIA" = "ICC/INDIA", "PAK" = "ICC/PAK", "SA" = "ICC/SA","WI" ="ICC/WI")
```

9. Create a new variable for the number of years the player has played cricket.

```
years = c()
for (year in str_match_all(origin_runs$Span,pattern = "\\d+")) {
years = c(years, as.numeric(year[2,])-as.numeric(year[1,]))
}
origin_runs$Years = years
origin_runs %>% head()
```

```
##   X        Player      Span Mat Inns NO  Runs  HS   Ave Centuries Fifties Zeros
## 1 1  SR Tendulkar 1989-2013 200  329 33 15921 248 53.78        51      68    14
## 2 2    RT Ponting 1995-2012 168  287 29 13378 257 51.85        41      62    17
## 3 3     JH Kallis 1995-2013 166  280 40 13289 224 55.37        45      58    16
## 4 4      R Dravid 1996-2012 164  286 32 13288 270 52.31        36      63     8
## 5 5       AN Cook 2006-2018 161  291 16 12472 294 45.35        33      57     9
## 6 6 KC Sangakkara 2000-2015 134  233 17 12400 319 57.40        38      52    11
##   Country Years
## 1   INDIA    24
## 2     AUS    17
## 3      SA    18
## 4   INDIA    16
## 5     ENG    12
## 6      SL    15
```

10. Save the dataset

```
write.csv(origin_runs, "runs.csv",row.names = FALSE )
```

11.Who scored highest test score? and from which country?

```
origin_runs[origin_runs$HS %>% which.max(),]$Player
```

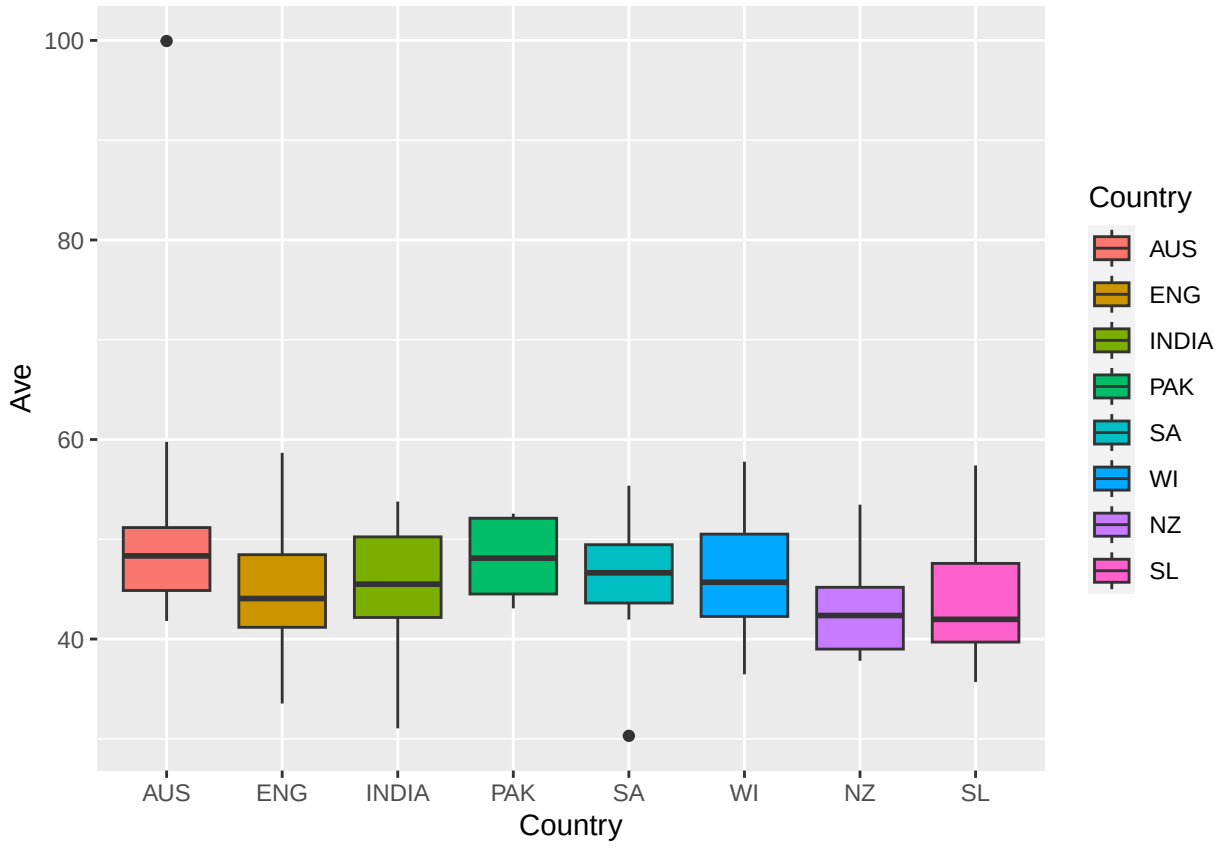
```
## [1] "BC Lara"
```

```
origin_runs[origin_runs$HS %>% which.max(),]$Country
```

```
## [1] WI
## Levels: AUS ENG INDIA PAK SA WI NZ SL
```

12.Compare the batting averages from each country

```
origin_runs %>% group_by(Country) %>%
  ggplot(aes(y = Ave, x = Country,fill = Country)) + geom_boxplot()
```



13. Create a bar graph for the proportion of centuries from each country?

```
#origin_runs %>% group_by(Country) %>%
#  ggplot(aes(y = Centuries, x = Country)) + geom_bar()
```

14. Build up a model to predict the average score by using the following predicters.

Mat, NO, HS, Centuries, Fifties, Zeros, Country, Years

```
origin_runs
```

```
##      X           Player      Span Mat Inns NO  Runs  HS   Ave Centuries Fifties
## 1    1     SR Tendulkar 1989-2013 200  329 33 15921 248 53.78        51      68
## 2    2       RT Ponting 1995-2012 168  287 29 13378 257 51.85        41      62
## 3    3        JH Kallis 1995-2013 166  280 40 13289 224 55.37        45      58
## 4    4         R Dravid 1996-2012 164  286 32 13288 270 52.31        36      63
## 5    5          AN Cook 2006-2018 161  291 16 12472 294 45.35        33      57
## 6    6    KC Sangakkara 2000-2015 134  233 17 12400 319 57.40        38      52
## 7    7          BC Lara 1990-2006 131  232  6 11953 400 52.88        34      48
## 8    8    S Chanderpaul 1994-2015 164  280 49 11867 203 51.37        30      66
## 9    9 DPMD Jayawardene 1997-2014 149  252 15 11814 374 49.84        34      50
## 10 10        AR Border 1978-1994 156  265 44 11174 205 50.56        27      63
## 11 11         SR Waugh 1985-2004 168  260 46 10927 200 51.06        32      50
## 12 12      SM Gavaskar 1971-1987 125  214 16 10122 236 51.12        34      45
## 13 13      Younis Khan 2000-2017 118  213 19 10099 313 52.05        34      33
## 14 14          JE Root 2012-2022 117  216 15  9889 254 49.19        25      53
## 15 15          HM Amla 2004-2019 124  215 16  9282 311 46.64        28      41
## 16 16        GC Smith 2002-2014 117  205 13  9265 277 48.25        27      38
```

```
## 17 17         GA Gooch 1975-1995 118  215  6  8900 333 42.58        20       46
## 18 18    Javed Miandad 1976-1993 124  189 21  8832 280 52.57        23       43
## 19 19   Inzamam-ul-Haq 1992-2007 120  200 22  8830 329 49.60        25       46
## 20 20       VVS Laxman 1996-2012 134  225 34  8781 281 45.97        17       56
## 21 21   AB de Villiers 2004-2018 114  191 18  8765 278 50.66        22       46
## 22 22        MJ Clarke 2004-2015 115  198 22  8643 329 49.10        28       27
## 23 23        ML Hayden 1994-2009 103  184 14  8625 380 50.73        30       29
## 24 24         V Sehwag 2001-2013 104  180  6  8586 319 49.34        23       32
## 25 25     IVA Richards 1974-1991 121  182 12  8540 291 50.23        24       45
## 26 26       AJ Stewart 1990-2003 133  235 21  8463 190 39.54        15       45
## 27 27         DI Gower 1978-1992 117  204 18  8231 215 44.25        18       39
## 28 28     KP Pietersen 2005-2014 104  181  8  8181 227 47.28        23       35
## 29 29        G Boycott 1964-1982 108  193 23  8114 246 47.72        22       42
## 30 30          V Kohli 2011-2022 101  171 10  8043 254 49.95        27       28
## 31 31        GS Sobers 1954-1974  93  160 21  8032 365 57.78        26       30
## 32 32         ME Waugh 1991-2002 128  209 17  8029 153 41.81        20       47
## 33 33        SPD Smith 2010-2022  85  151 17  8010 239 59.77        27       36
## 34 34        DA Warner 2011-2022  94  172  7  7753 335 46.98        24       34
## 35 35      MA Atherton 1989-2001 115  212  7  7728 185 37.69        16       46
## 36 36          IR Bell 2004-2015 118  205 24  7727 235 42.69        22       46
## 37 37        JL Langer 1993-2007 105  182 12  7696 250 45.27        23       30
## 38 38      LRPL Taylor 2007-2022 112  196 24  7683 290 44.66        19       35
## 39 39       MC Cowdrey 1954-1975 114  188 15  7624 182 44.06        22       38
## 40 40     CG Greenidge 1974-1991 108  185 16  7558 226 44.72        19       34
## 41 41  Mohammad Yousuf 1998-2010  90  156 12  7530 223 52.29        24       33
## 42 42        MA Taylor 1989-1999 104  186 13  7525 334 43.49        19       40
## 43 43         CH Lloyd 1966-1985 110  175 14  7515 242 46.67        19       39
## 44 44        DL Haynes 1978-1994 116  202 25  7487 184 42.29        18       39
## 45 45          DC Boon 1984-1996 107  190 20  7422 200 43.65        21       32
## 46 46        G Kirsten 1993-2004 101  176 15  7289 275 45.27        21       34
## 47 47    KS Williamson 2010-2021  86  150 14  7272 251 53.47        24       33
## 48 48       WR Hammond 1927-1947  85  140 16  7249 336 58.45        22       24
## 49 49         CH Gayle 2000-2014 103  182 11  7214 333 42.18        15       37
## 50 50       SC Ganguly 1996-2008 113  188 17  7212 239 42.17        16       35
## 51 51       SP Fleming 1994-2008 111  189 10  7172 274 40.06         9       46
## 52 52      GS Chappell 1970-1984  87  151 19  7110 247 53.86        24       31
## 53 53       AJ Strauss 2004-2012 100  178  6  7037 177 40.91        21       27
## 54 54        Azhar Ali 2010-2022  94  174 11  7021 302 43.07        19       35
## 55 55       DG Bradman 1928-1948  52   80 10  6996 334 99.94        29       13
## 56 56    ST Jayasuriya 1991-2007 110  188 14  6973 340 40.07        14       31
## 57 57         L Hutton 1937-1955  79  138 15  6971 364 56.67        19       33
## 58 58    DB Vengsarkar 1976-1992 116  185 22  6868 166 42.13        17       35
## 59 59    KF Barrington 1955-1968  82  131 15  6806 256 58.67        20       35
## 60 60        GP Thorpe 1993-2005 100  179 28  6744 200 44.66        16       39
## 61 61        CA Pujara 2010-2022  95  162  9  6713 206 43.87        18       32
## 62 62      BB McCullum 2004-2016 101  176  9  6453 302 38.64        12       31
## 63 63       AD Mathews 2009-2022  94  169 23  6432 200 44.05        11       37
## 64 64      PA de Silva 1984-2002  93  159 11  6361 267 42.97        20       22
## 65 65       MEK Hussey 2005-2013  79  137 16  6235 195 51.52        19       29
## 66 66         RB Kanhai 1957-1974  79  137  6  6227 256 47.53        15       28
## 67 67     M Azharuddin 1984-2000  99  147  9  6215 199 45.03        22       21
## 68 68         HH Gibbs 1996-2008  90  154  7  6167 228 41.95        14       26
## 69 69        RN Harvey 1948-1963  79  137 10  6149 205 48.41        21       24
## 70 70     GR Viswanath 1969-1983  91  155 10  6080 222 41.93        14       35
```

```
## 71 71    RB Richardson 1983-1995  86  146 12  5949 194 44.39        16       27
## 72 72        RR Sarwan 2000-2011  87  154  8  5842 291 40.01        15       31
## 73 73   ME Trescothick 2000-2006  76  143 10  5825 219 43.79        14       29
## 74 74      DCS Compton 1937-1957  78  131 15  5807 278 50.06        17       28
## 75 75     Saleem Malik 1982-1999 103  154 22  5768 237 43.69        15       29
## 76 76        N Hussain 1990-2004  96  171 16  5764 207 37.18        14       33
## 77 77        CL Hooper 1987-2002 102  173 15  5762 233 36.46        13       27
## 78 78       MP Vaughan 1999-2008  82  147  9  5719 197 41.44        18       18
## 79 79  FDM Karunaratne 2012-2022  76  147  5  5620 244 39.57        14       27
## 80 80     AC Gilchrist 1999-2008  96  137 20  5570 204 47.60        17       26
## 81 81       MV Boucher 1997-2012 147  206 24  5515 125 30.30         5       35
## 82 82      MS Atapattu 1990-2007  90  156 15  5502 249 39.02        16       17
## 83 83       TM Dilshan 1999-2013  87  145 11  5492 193 40.98        16       23
## 84 84   TT Samaraweera 2001-2013  81  132 20  5462 231 48.76        14       30
## 85 85         MD Crowe 1982-1995  77  131 11  5444 299 45.36        17       18
## 86 86         JB Hobbs 1908-1930  61  102  7  5410 211 56.94        15       28
## 87 87       KD Walters 1965-1981  74  125 14  5357 250 48.26        15       33
## 88 88      IM Chappell 1964-1980  75  136 10  5345 196 42.42        14       26
## 89 89        JG Wright 1978-1993  82  148  7  5334 185 37.82        12       23
## 90 90        MJ Slater 1993-2001  74  131  7  5312 219 42.83        14       21
## 91 91      N Kapil Dev 1978-1994 131  184 15  5248 163 31.05         8       27
## 92 92         WM Lawry 1961-1971  67  123 12  5234 210 47.15        13       27
## 93 93    Misbah-ul-Haq 2001-2017  75  132 20  5222 161 46.62        10       39
## 94 94        IT Botham 1977-1992 102  161  6  5200 208 33.54        14       22
## 95 95        JH Edrich 1963-1976  77  127  9  5138 310 43.54        12       24
## 96 96      A Ranatunga 1982-2000  93  155 12  5105 135 35.69         4       38
## 97 97     Zaheer Abbas 1969-1985  78  124 11  5062 274 44.79        12       20
## 98 98        BA Stokes 2013-2022  79  146  5  5061 258 35.89        11       26
##    Zeros Country Years
## 1     14   INDIA    24
## 2     17     AUS    17
## 3     16      SA    18
## 4      8   INDIA    16
## 5      9     ENG    12
## 6     11      SL    15
## 7     17      WI    16
## 8     15      WI    21
## 9     15      SL    17
## 10    11     AUS    16
## 11    22     AUS    19
## 12    12   INDIA    16
## 13    19     PAK    17
## 14    11     ENG    10
## 15    13      SA    15
## 16    11      SA    12
## 17    13     ENG    20
## 18     6     PAK    17
## 19    15     PAK    15
## 20    14   INDIA    16
## 21     8      SA    14
## 22     9     AUS    11
## 23    14     AUS    15
## 24    16   INDIA    12
## 25    10      WI    17
```

```
## 26    14     ENG    13
## 27     7     ENG    14
## 28    10     ENG     9
## 29    10     ENG    18
## 30    14   INDIA    11
## 31    12      WI    20
## 32    19     AUS    11
## 33     6     AUS    12
## 34    11     AUS    11
## 35    20     ENG    12
## 36    14     ENG    11
## 37    11     AUS    14
## 38    14      NZ    15
## 39     9     ENG    21
## 40    11      WI    17
## 41    11     PAK    12
## 42     5     AUS    10
## 43     4      WI    19
## 44    10      WI    16
## 45    16     AUS    12
## 46    13      SA    11
## 47     9      NZ    11
## 48     4     ENG    20
## 49    15      WI    14
## 50    13   INDIA    12
## 51    16      NZ    14
## 52    12     AUS    14
## 53    15     ENG     8
## 54    18     PAK    12
## 55     7     AUS    20
## 56    15      SL    16
## 57     5     ENG    18
## 58    15   INDIA    16
## 59     5     ENG    13
## 60    12     ENG    12
## 61    11   INDIA    12
## 62    14      NZ    12
## 63     2      SL    13
## 64     7      SL    18
## 65    12     AUS     8
## 66     7      WI    17
## 67     5   INDIA    16
## 68    11      SA    12
## 69     7     AUS    15
## 70    10   INDIA    14
## 71     8      WI    12
## 72    12      WI    11
## 73    12     ENG     6
## 74    10     ENG    20
## 75    12     PAK    17
## 76    14     ENG    14
## 77    13      WI    15
## 78     9     ENG     9
## 79    12      SL    10
```

```
## 80    14     AUS     9
## 81    17      SA    15
## 82    22      SL    17
## 83    14      SL    14
## 84    11      SL    12
## 85     9      NZ    13
## 86     4     ENG    22
## 87     4     AUS    16
## 88    11     AUS    16
## 89     7      NZ    15
## 90     9     AUS     8
## 91    16   INDIA    16
## 92     6     AUS    10
## 93     9     PAK    16
## 94    14     ENG    15
## 95     6     ENG    13
## 96    12      SL    18
## 97    10     PAK    16
## 98    13     ENG     9
```

```
lm <-lm( Ave ~ Mat + NO + HS + Centuries + Fifties + Zeros + Country + Years, data = origin_runs)
summary(lm)
```

```
##
## Call:
## lm(formula = Ave ~ Mat + NO + HS + Centuries + Fifties + Zeros +
##     Country + Years, data = origin_runs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1215 -2.4643 -0.2621  1.8771 29.6263
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  43.003372   3.278618  13.116  < 2e-16 ***
## Mat          -0.296627   0.043101  -6.882 1.04e-09 ***
## NO            0.273374   0.082040   3.332  0.00129 **
## HS            0.019197   0.009758   1.967  0.05250 .
## Centuries     0.779806   0.087483   8.914 9.72e-14 ***
## Fifties       0.157642   0.080894   1.949  0.05470 .
## Zeros        -0.160697   0.135962  -1.182  0.24061
## CountryENG   -1.760646   1.495809  -1.177  0.24254
## CountryINDIA -0.462641   1.794038  -0.258  0.79714
## CountryPAK   -1.906002   1.960429  -0.972  0.33376
## CountrySA    -0.245212   2.068964  -0.119  0.90594
## CountryWI    -2.235681   1.807769  -1.237  0.21968
## CountryNZ    -1.904071   2.215692  -0.859  0.39262
## CountrySL    -2.969234   1.858457  -1.598  0.11391
## Years         0.502533   0.156083   3.220  0.00183 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.503 on 83 degrees of freedom
## Multiple R-squared:  0.7388, Adjusted R-squared:  0.6947
## F-statistic: 16.77 on 14 and 83 DF,  p-value: < 2.2e-16
```

15. Check the assumptions

Reject the alternative assumption, they have linear relationships because p value is small. But we can only trust that "Mat","NO","Centuries" and "Years" have statistical significant.

16. Predict the batting average of a player with the following statistics?

```
predict_data <- data.frame(
  Mat = 85, NO = 17, HS = mean(origin_runs$HS),
  Centuries = 27, Fifties = 36, Zeros = mean(origin_runs$Zeros),
  Country = "AUS",
  Years = 12
)
predict(lm, newdata = predict_data)
```

```
##        1
## 58.16065
```

17. Can we use this model to predict the batting average of any player?

No, this model is not credible.