A2-R.R

ujjwa

2025-06-19

```r
# SCMA 632 Assignment - Final R Script
# Objective: IPL Player Performance and Salary Analysis using R

# --- Project Setup and Package Management ---

# This section sets up your R environment and installs/loads the necessary packages

# Set the base directory for the project
BASE <- "C:\\Users\\ujjwa\\Documents\\VCU\\Pre-Course\\SCMA632\\Assignments\\A2\\R"
setwd(BASE)  # Change working directory to where your data files are located
getwd()  # Confirm working directory
```

```
## [1] "C:/Users/ujjwa/Documents/VCU/Pre-Course/SCMA632/Assignments/A2/R"
```

```r
# Define a helper function to install packages if not already installed
install <- function(pkg) {
  if (!require(pkg, character.only = TRUE)) {
    install.packages(pkg, dependencies = TRUE, quiet = TRUE)
  }
}

# Define a helper function to load packages
load <- function(pkg) {
  library(pkg, character.only = TRUE, quietly = TRUE)
}

# Required packages for this analysis
pkgs <- c("dplyr", "readr", "readxl", "lubridate", "stringdist", "stats", "fitdistrplus")
lapply(pkgs, install)
```

```
## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Loading required package: readr

## Loading required package: readxl

## Loading required package: lubridate

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union

## Loading required package: stringdist

## Loading required package: fitdistrplus

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## Loading required package: survival

## [[1]]
## NULL
##
## [[2]]
## NULL
##
## [[3]]
## NULL
##
## [[4]]
## NULL
##
## [[5]]
## NULL
##
## [[6]]
## NULL
##
## [[7]]
## NULL
```

```r
lapply(pkgs, load)
```

```
## [[1]]
##  [1] "fitdistrplus" "survival"     "MASS"         "stringdist"   "lubridate"
##  [6] "readxl"       "readr"        "dplyr"        "stats"        "graphics"
## [11] "grDevices"    "utils"        "datasets"     "methods"      "base"
##
## [[2]]
##  [1] "fitdistrplus" "survival"     "MASS"         "stringdist"   "lubridate"
##  [6] "readxl"       "readr"        "dplyr"        "stats"        "graphics"
## [11] "grDevices"    "utils"        "datasets"     "methods"      "base"
##
## [[3]]
##  [1] "fitdistrplus" "survival"     "MASS"         "stringdist"   "lubridate"
##  [6] "readxl"       "readr"        "dplyr"        "stats"        "graphics"
## [11] "grDevices"    "utils"        "datasets"     "methods"      "base"
##
## [[4]]
##  [1] "fitdistrplus" "survival"     "MASS"         "stringdist"   "lubridate"
##  [6] "readxl"       "readr"        "dplyr"        "stats"        "graphics"
## [11] "grDevices"    "utils"        "datasets"     "methods"      "base"
##
## [[5]]
##  [1] "fitdistrplus" "survival"     "MASS"         "stringdist"   "lubridate"
##  [6] "readxl"       "readr"        "dplyr"        "stats"        "graphics"
## [11] "grDevices"    "utils"        "datasets"     "methods"      "base"
##
## [[6]]
##  [1] "fitdistrplus" "survival"     "MASS"         "stringdist"   "lubridate"
##  [6] "readxl"       "readr"        "dplyr"        "stats"        "graphics"
## [11] "grDevices"    "utils"        "datasets"     "methods"      "base"
##
## [[7]]
##  [1] "fitdistrplus" "survival"     "MASS"         "stringdist"   "lubridate"
##  [6] "readxl"       "readr"        "dplyr"        "stats"        "graphics"
## [11] "grDevices"    "utils"        "datasets"     "methods"      "base"
```

```r
# --- Data Import ---

# Read in the IPL performance and salary datasets

# Define file paths
perf_path <- "datasets/IPL_ball_by_ball_updated till 2024.csv"
salary_path <- "datasets/IPL SALARIES 2024.xlsx"

# Read the datasets
df <- read_csv(perf_path)
```

```
## Rows: 255759 Columns: 19
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
```

```
## chr (12): Date, Season, Batting team, Bowling team, Bowler, Striker, Non Str...
## dbl  (7): Match id, Innings No, Ball No, runs_scored, extras, score, wicket_...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
salary_df <- read_excel(salary_path)

# --- Data Preprocessing ---

# Select only relevant columns and extract the year from the date column
columns_to_select <- c("Match id", "Date", "Season", "Innings No",
                       "Bowler", "Striker", "runs_scored", "wicket_confirmation")
df <- df %>% dplyr::select(all_of(columns_to_select))
df$Year <- year(dmy(df$Date))  # Extract year from the date

# --- Aggregate Data ---

# Summarize total runs by each batsman
runs <- df %>%
  group_by(Year, `Innings No`, Striker) %>%
  summarise(runs_scored = sum(runs_scored), .groups = 'drop')

# Summarize total wickets by each bowler
wickets <- df %>%
  group_by(Year, `Innings No`, Bowler) %>%
  summarise(wicket_confirmation = sum(wicket_confirmation), .groups = 'drop')

# --- Identify Top 3 Performers Each Year ---

years <- unique(runs$Year)
for (yr in years) {
  cat("Year:", yr, "\n\nTop 3 Run Scorers:\n")
  print(runs %>% filter(Year == yr) %>%
          group_by(Striker) %>%
          summarise(runs = sum(runs_scored), .groups = 'drop') %>%
          arrange(desc(runs)) %>% head(3))

  cat("\nTop 3 Wicket Takers:\n")
  print(wickets %>% filter(Year == yr) %>%
          group_by(Bowler) %>%
          summarise(wickets = sum(wicket_confirmation), .groups = 'drop') %>%
          arrange(desc(wickets)) %>% head(3))

  cat("\n", strrep("=", 50), "\n\n")
}
```

```
## Year: 2008
##
## Top 3 Run Scorers:
## # A tibble: 3 x 2
##   Striker        runs
##   <chr>         <dbl>
## 1 SE Marsh        616
```

```
## 2 G Gambhir       534
## 3 ST Jayasuriya   514
##
## Top 3 Wicket Takers:
## # A tibble: 3 x 2
##   Bowler        wickets
##   <chr>           <dbl>
## 1 Sohail Tanvir      24
## 2 IK Pathan          20
## 3 JA Morkel          20
##
##  ================================================
##
## Year: 2009
##
## Top 3 Run Scorers:
## # A tibble: 3 x 2
##   Striker         runs
##   <chr>          <dbl>
## 1 ML Hayden        572
## 2 AC Gilchrist     495
## 3 AB de Villiers   465
##
## Top 3 Wicket Takers:
## # A tibble: 3 x 2
##   Bowler   wickets
##   <chr>      <dbl>
## 1 RP Singh      26
## 2 A Kumble      22
## 3 A Nehra       22
##
##  ================================================
##
## Year: 2010
##
## Top 3 Run Scorers:
## # A tibble: 3 x 2
##   Striker       runs
##   <chr>        <dbl>
## 1 SR Tendulkar   618
## 2 JH Kallis      572
## 3 SK Raina       528
##
## Top 3 Wicket Takers:
## # A tibble: 3 x 2
##   Bowler          wickets
##   <chr>             <dbl>
## 1 PP Ojha              22
## 2 A Mishra             20
## 3 Harbhajan Singh      20
##
##  ================================================
##
## Year: 2011
```

```
##
## Top 3 Run Scorers:
## # A tibble: 3 x 2
##   Striker       runs
##   <chr>        <dbl>
## 1 CH Gayle       608
## 2 V Kohli        557
## 3 SR Tendulkar   553
##
## Top 3 Wicket Takers:
## # A tibble: 3 x 2
##   Bowler     wickets
##   <chr>        <dbl>
## 1 SL Malinga      30
## 2 MM Patel        22
## 3 S Aravind       22
##
##  ================================================
##
## Year: 2012
##
## Top 3 Run Scorers:
## # A tibble: 3 x 2
##   Striker    runs
##   <chr>     <dbl>
## 1 CH Gayle    733
## 2 G Gambhir   590
## 3 S Dhawan    569
##
## Top 3 Wicket Takers:
## # A tibble: 3 x 2
##   Bowler     wickets
##   <chr>        <dbl>
## 1 M Morkel        30
## 2 SP Narine       29
## 3 SL Malinga      25
##
##  ================================================
##
## Year: 2013
##
## Top 3 Run Scorers:
## # A tibble: 3 x 2
##   Striker     runs
##   <chr>      <dbl>
## 1 MEK Hussey   733
## 2 CH Gayle     720
## 3 V Kohli      639
##
## Top 3 Wicket Takers:
## # A tibble: 3 x 2
##   Bowler        wickets
##   <chr>           <dbl>
## 1 DJ Bravo           34
```

```
## 2 JP Faulkner        33
## 3 R Vinay Kumar      27
##
##  ================================================
##
## Year: 2014
##
## Top 3 Run Scorers:
## # A tibble: 3 x 2
##   Striker     runs
##   <chr>      <dbl>
## 1 RV Uthappa   660
## 2 DR Smith     566
## 3 GJ Maxwell   552
##
## Top 3 Wicket Takers:
## # A tibble: 3 x 2
##   Bowler    wickets
##   <chr>       <dbl>
## 1 MM Sharma      26
## 2 SP Narine      22
## 3 B Kumar        21
##
##  ================================================
##
## Year: 2015
##
## Top 3 Run Scorers:
## # A tibble: 3 x 2
##   Striker      runs
##   <chr>       <dbl>
## 1 DA Warner     562
## 2 AM Rahane     540
## 3 LMP Simmons   540
##
## Top 3 Wicket Takers:
## # A tibble: 3 x 2
##   Bowler     wickets
##   <chr>        <dbl>
## 1 DJ Bravo        28
## 2 SL Malinga      26
## 3 A Nehra         25
##
##  ================================================
##
## Year: 2016
##
## Top 3 Run Scorers:
## # A tibble: 3 x 2
##   Striker         runs
##   <chr>          <dbl>
## 1 V Kohli          973
## 2 DA Warner        848
## 3 AB de Villiers   687
```

```
##
## Top 3 Wicket Takers:
## # A tibble: 3 x 2
##   Bowler    wickets
##   <chr>       <dbl>
## 1 B Kumar        24
## 2 SR Watson      23
## 3 YS Chahal      22
##
##  ================================================
##
## Year: 2017
##
## Top 3 Run Scorers:
## # A tibble: 3 x 2
##   Striker    runs
##   <chr>     <dbl>
## 1 DA Warner   641
## 2 G Gambhir   498
## 3 S Dhawan    479
##
## Top 3 Wicket Takers:
## # A tibble: 3 x 2
##   Bowler     wickets
##   <chr>        <dbl>
## 1 B Kumar         28
## 2 JD Unadkat      27
## 3 JJ Bumrah       23
##
##  ================================================
##
## Year: 2018
##
## Top 3 Run Scorers:
## # A tibble: 3 x 2
##   Striker        runs
##   <chr>         <dbl>
## 1 KS Williamson   735
## 2 RR Pant         684
## 3 KL Rahul        659
##
## Top 3 Wicket Takers:
## # A tibble: 3 x 2
##   Bowler      wickets
##   <chr>         <dbl>
## 1 AJ Tye           28
## 2 S Kaul           24
## 3 Rashid Khan      23
##
##  ================================================
##
## Year: 2019
##
## Top 3 Run Scorers:
```

```
## # A tibble: 3 x 2
##   Striker    runs
##   <chr>     <dbl>
## 1 DA Warner   692
## 2 KL Rahul    593
## 3 Q de Kock   529
##
## Top 3 Wicket Takers:
## # A tibble: 3 x 2
##   Bowler      wickets
##   <chr>         <dbl>
## 1 K Rabada         29
## 2 Imran Tahir      26
## 3 JJ Bumrah        23
##
##  ================================================
##
## Year: 2020
##
## Top 3 Run Scorers:
## # A tibble: 3 x 2
##   Striker    runs
##   <chr>     <dbl>
## 1 KL Rahul    676
## 2 S Dhawan    618
## 3 DA Warner   548
##
## Top 3 Wicket Takers:
## # A tibble: 3 x 2
##   Bowler    wickets
##   <chr>       <dbl>
## 1 K Rabada       32
## 2 JJ Bumrah      30
## 3 TA Boult       26
##
##  ================================================
##
## Year: 2021
##
## Top 3 Run Scorers:
## # A tibble: 3 x 2
##   Striker       runs
##   <chr>        <dbl>
## 1 RD Gaikwad     635
## 2 F du Plessis   633
## 3 KL Rahul       626
##
## Top 3 Wicket Takers:
## # A tibble: 3 x 2
##   Bowler     wickets
##   <chr>        <dbl>
## 1 HV Patel        35
## 2 Avesh Khan      27
## 3 JJ Bumrah       22
```

```
##
##  ================================================
##
## Year: 2022
##
## Top 3 Run Scorers:
## # A tibble: 3 x 2
##   Striker     runs
##   <chr>      <dbl>
## 1 JC Buttler   863
## 2 KL Rahul     616
## 3 Q de Kock    508
##
## Top 3 Wicket Takers:
## # A tibble: 3 x 2
##   Bowler       wickets
##   <chr>          <dbl>
## 1 YS Chahal         29
## 2 PWH de Silva      27
## 3 K Rabada          23
##
##  ================================================
##
## Year: 2023
##
## Top 3 Run Scorers:
## # A tibble: 3 x 2
##   Striker       runs
##   <chr>        <dbl>
## 1 Shubman Gill   890
## 2 F du Plessis   730
## 3 DP Conway      672
##
## Top 3 Wicket Takers:
## # A tibble: 3 x 2
##   Bowler         wickets
##   <chr>            <dbl>
## 1 MM Sharma           31
## 2 Mohammed Shami      28
## 3 Rashid Khan         28
##
##  ================================================
##
## Year: 2024
##
## Top 3 Run Scorers:
## # A tibble: 3 x 2
##   Striker          runs
##   <chr>           <dbl>
## 1 RD Gaikwad        509
## 2 V Kohli           500
## 3 B Sai Sudharsan   418
##
## Top 3 Wicket Takers:
```

```
## # A tibble: 3 x 2
##   Bowler         wickets
##   <chr>            <dbl>
## 1 HV Patel            19
## 2 Mukesh Kumar        15
## 3 Arshdeep Singh      14
##
##  ================================================
```

```r
# --- Name Matching Between Datasets (Fuzzy Matching) ---

match_names <- function(name, choices, threshold = 0.2) {
  if (is.na(name)) return(NA)
  dists <- stringdist(name, choices, method = "jw")
  min_dist <- min(dists)
  if (min_dist <= threshold) return(choices[which.min(dists)])
  return(NA)
}

# --- Correlation Between Salary and Performance (2024) ---

runs_2024 <- df %>%
  filter(Year == 2024) %>%
  group_by(Striker) %>%
  summarise(runs_scored = sum(runs_scored), .groups = 'drop')

wickets_2024 <- df %>%
  filter(Year == 2024) %>%
  group_by(Bowler) %>%
  summarise(wicket_confirmation = sum(wicket_confirmation), .groups = 'drop')

salary_df$Matched_Striker <- sapply(salary_df$Player, match_names, choices = runs_2024$Striker)
salary_df$Matched_Bowler <- sapply(salary_df$Player, match_names, choices = wickets_2024$Bowler)

striker_merged <- merge(salary_df, runs_2024, by.x = "Matched_Striker", by.y = "Striker")
bowler_merged <- merge(salary_df, wickets_2024, by.x = "Matched_Bowler", by.y = "Bowler")

cor_striker <- cor(striker_merged$Rs, striker_merged$runs_scored, use = "complete.obs")
cor_bowler <- cor(bowler_merged$Rs, bowler_merged$wicket_confirmation, use = "complete.obs")

cat("\nCorrelation between Salary and Runs in 2024:", cor_striker)
```

```
##
## Correlation between Salary and Runs in 2024: 0.4531945
```

```r
cat("\nCorrelation between Salary and Wickets in 2024:", cor_bowler)
```

```
##
## Correlation between Salary and Wickets in 2024: 0.2137848
```

```r
# --- Distribution Fitting for Assigned Player: N Pooran ---
```

```r
n_pooran_runs <- df %>%
  group_by(Year, Striker, `Match id`) %>%
  summarise(runs_scored = sum(runs_scored), .groups = 'drop') %>%
  filter(Striker == "N Pooran") %>%
  pull(runs_scored)

n_pooran_runs_pos <- n_pooran_runs[n_pooran_runs > 0]

fit_norm <- fitdist(n_pooran_runs, "norm")
fit_gamma <- fitdist(n_pooran_runs, "gamma")
fit_exp <- fitdist(n_pooran_runs, "exp")
fit_lnorm <- fitdist(n_pooran_runs_pos, "lnorm")

gof <- gofstat(list(fit_norm, fit_gamma, fit_exp))
gof_lnorm <- if (length(n_pooran_runs_pos) == length(n_pooran_runs)) gofstat(list(fit_lnorm)) else NULL

print(gof)
```

```
## Goodness-of-fit statistics
##                               1-mle-norm 2-mle-gamma 3-mle-exp
## Kolmogorov-Smirnov statistic  0.1371003   0.1734986 0.1458793
## Cramer-von Mises statistic    0.2659139   0.4227711 0.2878949
## Anderson-Darling statistic    1.7800216         Inf       Inf
##
## Goodness-of-fit criteria
##                                 1-mle-norm 2-mle-gamma 3-mle-exp
## Akaike's Information Criterion   605.0947    513.6439  565.3804
## Bayesian Information Criterion   609.5337    518.0829  567.5999
```

```r
cat("\nLognormal fit (positive data only):\n")
```

```
##
## Lognormal fit (positive data only):
```

```r
print(gof_lnorm)
```

```
## NULL
```

```r
ks <- ks.test(n_pooran_runs, "pgamma", shape = fit_gamma$estimate["shape"], rate = fit_gamma$estimate["
```

```
## Warning in ks.test.default(n_pooran_runs, "pgamma", shape =
## fit_gamma$estimate["shape"], : ties should not be present for the one-sample
## Kolmogorov-Smirnov test
```

```r
print(ks)
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  n_pooran_runs
## D = 0.1735, p-value = 0.03335
## alternative hypothesis: two-sided
```

```r
# --- Distribution Fitting for Top 3 Batsmen and Bowlers (Last 3 Seasons) ---

get_best_fit <- function(data) {
  if (length(data) < 2) return(list(best_dist = "Too few data", gof = NULL))
  data_pos <- data[data > 0]
  use_pos <- all(data > 0)
  data_used <- if (use_pos) data else data_pos

  fits <- list(
    norm = tryCatch(fitdist(data_used, "norm"), error = function(e) NULL),
    gamma = tryCatch(fitdist(data_used, "gamma"), error = function(e) NULL),
    exp = tryCatch(fitdist(data_used, "exp"), error = function(e) NULL)
  )
  if (all(data_used > 0)) {
    fits$lnorm <- tryCatch(fitdist(data_used, "lnorm"), error = function(e) NULL)
  }

  fits <- Filter(Negate(is.null), fits)

  if (length(fits) <= 1) return(list(best_dist = names(fits), gof = NULL))

  lengths <- sapply(fits, function(f) length(f$data))
  mode_len <- as.numeric(names(which.max(table(lengths))))
  fits_same <- fits[lengths == mode_len]

  if (length(fits_same) > 1) {
    gof <- tryCatch(gofstat(fits_same), error = function(e) NULL)
    if (!is.null(gof)) {
      best <- names(which.max(gof$ks))
      return(list(best_dist = best, gof = gof))
    }
  }

  return(list(best_dist = "N/A", gof = NULL))
}

last_3_seasons <- sort(unique(df$Year), decreasing = TRUE)[1:3]

df_runs <- df %>%
  group_by(Year, Striker, `Match id`) %>%
  summarise(runs_scored = sum(runs_scored), .groups = 'drop')

df_wickets <- df %>%
  group_by(Year, Bowler, `Match id`) %>%
  summarise(wicket_confirmation = sum(wicket_confirmation), .groups = 'drop')

for (year in last_3_seasons) {
  cat("\n===== Year:", year, "=====")

  top_batsmen <- df_runs %>%
    filter(Year == year) %>%
    group_by(Striker) %>%
    summarise(total_runs = sum(runs_scored), .groups = 'drop') %>%
```

```
    arrange(desc(total_runs)) %>%
    slice(1:3) %>% pull(Striker)

  top_bowlers <- df_wickets %>%
    filter(Year == year) %>%
    group_by(Bowler) %>%
    summarise(total_wickets = sum(wicket_confirmation), .groups = 'drop') %>%
    arrange(desc(total_wickets)) %>%
    slice(1:3) %>% pull(Bowler)

  cat("\nTop 3 Batsmen Distribution Fits:\n")
  for (batsman in top_batsmen) {
    player_data <- df_runs %>% filter(Year == year, Striker == batsman) %>% pull(runs_scored)
    result <- get_best_fit(player_data)
    cat("-", batsman, ": Best Fit:", result$best_dist, "\n")
  }

  cat("\nTop 3 Bowlers Distribution Fits:\n")
  for (bowler in top_bowlers) {
    player_data <- df_wickets %>% filter(Year == year, Bowler == bowler) %>% pull(wicket_confirmation)
    result <- get_best_fit(player_data)
    cat("-", bowler, ": Best Fit:", result$best_dist, "\n")
  }
}
```

```
##
## ===== Year: 2024 =====
## Top 3 Batsmen Distribution Fits:
## - RD Gaikwad : Best Fit: 4-mle-lnorm
## - V Kohli : Best Fit: 3-mle-exp
## - B Sai Sudharsan : Best Fit: 3-mle-exp
##
## Top 3 Bowlers Distribution Fits:
## - HV Patel : Best Fit: N/A
## - Mukesh Kumar : Best Fit: N/A
## - Arshdeep Singh : Best Fit: N/A
##
## ===== Year: 2023 =====
## Top 3 Batsmen Distribution Fits:
## - Shubman Gill : Best Fit: 3-mle-exp
## - F du Plessis : Best Fit: 3-mle-exp
## - DP Conway : Best Fit: 4-mle-lnorm
##
## Top 3 Bowlers Distribution Fits:
## - MM Sharma : Best Fit: 3-mle-exp
## - Mohammed Shami : Best Fit: N/A
## - Rashid Khan : Best Fit: N/A
##
## ===== Year: 2022 =====
## Top 3 Batsmen Distribution Fits:
## - JC Buttler : Best Fit: 4-mle-lnorm
## - KL Rahul : Best Fit: 4-mle-lnorm
## - Q de Kock : Best Fit: 1-mle-norm
```

```
##
## Top 3 Bowlers Distribution Fits:
## - YS Chahal : Best Fit: N/A
## - PWH de Silva : Best Fit: N/A
## - K Rabada : Best Fit: N/A
```