

Dear students,

Below given are the Assignment #1 submission guidelines of each team:

Due Date: 04:00PM of 23.03.2024

\*If necessary wherever required add more fields

|  |  |
| --- | --- |
| Title of the Minor Project: |  |
| Name of Team Player – 1 | Jangam Monesh |
| Roll No of Team Player - 1 | 1601-21-771-039 |
| Name of Team Player – 2 | Md.Mushtaq |
| Roll No of Team Player - 2 | 1601-21-771-050 |
| Name of Team Player – 3 | M.GopiPrashanthRaju |
| Roll No of Team Player - 3 | 1601-21-771-051 |
| Title of the Data Set | Cricksheet\_json\_files |
| **Description of Dataset:**  Size of the dataset  URL of the dataset  Summary / Description and Domain of each attribute  No. of missing values in each attribute | **SIZE**  83 json files containing meta info and ball to ball information  **URL:**  https://cricsheet.org/downloads/  **Summary:**  Meta Information:  Data Version: Version of the dataset.  Created: Date when the dataset was created.  Revision: Revision number of the dataset.  Match Information:  Date: Date of the match.  Event: Event name, match number, and group.  Gender: Gender of players (male).  Match Type: Type of match (ODI).  Match Type Number: Unique identifier for the match type.  City: City where the match took place.  Venue: Venue where the match was held.  Season: Season of the match.  Teams: Names of the teams playing the match.  Team Type: Type of teams (international).  Balls per Over: Number of balls per over.  Overs: Total number of overs per innings.  Toss: Toss decision and winner.  Officials:  Match Referees: Referees for the match.  TV Umpires: TV umpires for the match.  Umpires: On-field umpires for the match.  Outcome:  Winner: Winning team.  By: Margin of victory (wickets, runs, etc.).  Player Information:  Players: List of players for each team.  Innings:  Team: Name of the team batting.  Overs: Detailed information about each over, including deliveries.  **No. of missing values in each attribute**  No Missing Values |
|  |  |
| **Please mention if you have created the dataset or added some more fields to the dataset.**  Give detailed description of your contribution with resources from which the data is curated | **Contribution Description:**  We compiled the dataset by extracting data from CSV files in a ball-by-ball manner, capturing crucial details such as runs scored by batsmen, balls played, runs given by bowlers, balls bowled, and extras awarded. Additionally, we enhanced the dataset by including fields related to the performance of both batsmen and bowlers in each match, thereby providing comprehensive ball-by-ball statistics for analysis.  **Resources Used:**  CSV Files: We utilized CSV files containing ball-by-ball information for each of the 83 cricket matches between India and Australia from 2003 to 2023.  Cricket Match Archives: Our dataset is grounded in real cricket matches between India and Australia, sourced from match archives and online scorecards spanning two decades.  Data Extraction Tools: Leveraging scripting or programming languages such as Python, we meticulously extracted and processed the data from the CSV files, ensuring accuracy and completeness.  Cricket Statistics Databases: We may have consulted cricket statistics databases or APIs to cross-reference and validate the information, maintaining the dataset's reliability.  Overall, our collaborative effort involved gathering, processing, and structuring the ball-by-ball data to create an exhaustive dataset suitable for in-depth cricket match analysis and research. |
|  |  |
| **Description of Task#1** | Ball to Ball data conversion from the existing json file |
| Required Pre-processing | Conversion from json to csv |
| Attributes involved | Batter,Baller,runsperball,totalscore,extras,wickets,win margin,winner,tos winner,decision |
| R-Code with necessary comments | # Load required library  library(jsonlite)  # install.packages(hash)  jsonData <- read\_json(path = "/Users/morampudigopiprashanthraju/Desktop/DataScience/Minor\_Project-II/Data/IND vs AUS/ODI/2003/65244.json")  print(jsonData)  #Mushtaq's json path  # file\_path <- "D:/Minor\_Project-II/Data/IND vs AUS/ODI/2003/65244.json"  #Defining function  convert\_to\_csv <- function(file\_path){  jsonData <- read\_json(path = file\_path)  balls <- 0  flag <- TRUE  batter\_vector<-c()  baller\_vector <- c()  balls\_vector <- c()  runsperball\_vector <- c()  extra\_vector <- c()  country\_vector <- c()  Innings\_vector <- c()  toss <- c()  toss\_winner <- c()  Innings <- c(1, 2)  toss\_decision <- c()  match\_winner <- c()  margin <- c()  total\_score=c()  wickets=c()  cum\_wickets=0  for (inn in Innings) {  cummulative\_score=0  flag <- TRUE  balls<-0  cumm\_wickets=0  while (balls <= 300 && flag) {  # print(toss\_decision)  tryCatch({  over\_index <- balls %/% 6 + 1  if (over\_index > length(jsonData$innings[[inn]]$overs)) {  flag <- FALSE  break  }  over\_balls <- length(jsonData$innings[[inn]]$overs[[over\_index]]$deliveries)  currball <- 1  extras\_count=0  while (currball <= over\_balls) {  ball\_index <- (balls %/% 6) \* 6 + currball # Calculate the current ball index  toss <- c(toss, jsonData$info$toss$winner)  match\_winner <- c(match\_winner, jsonData$info$outcome$winner)  batter <- jsonData$innings[[inn]]$overs[[over\_index]]$deliveries[[currball]]$batter  bowler <- jsonData$innings[[inn]]$overs[[over\_index]]$deliveries[[currball]]$bowler  runs <- jsonData$innings[[inn]]$overs[[over\_index]]$deliveries[[currball]]$runs$batter  team <- jsonData$innings[[inn]]$team  batter\_vector<-c(batter\_vector,batter)  baller\_vector <- c(baller\_vector, bowler)  balls\_vector <- c(balls\_vector, balls%/%6 + 0.1\*(currball-extras\_count))  cummulative\_score=cummulative\_score+jsonData$innings[[inn]]$overs[[over\_index]]$deliveries[[currball]]$runs$total  runsperball\_vector <- c(runsperball\_vector, runs)  country\_vector <- c(country\_vector, team)  Innings\_vector <- c(Innings\_vector, inn)  toss\_winner<-c(toss\_winner,toss)  toss\_decision <- c(toss\_decision, jsonData$info$toss$decision)  if (!is.null(jsonData$innings[[inn]]$overs[[over\_index]]$deliveries[[currball]]$extras)) {  if(is.null(jsonData$innings[[inn]]$overs[[over\_index]]$deliveries[[currball]]$extras$legbyes))  {  extras\_count=extras\_count+1  }  extra\_vector <- c(extra\_vector, TRUE)  } else {  extra\_vector <- c(extra\_vector, FALSE)  }  if (!is.null(jsonData$info$outcome$by$wickets)) {  margin <- c(margin, paste(jsonData$info$outcome$by$wickets, " wickets"))  } else {  margin <- c(margin, paste(jsonData$info$outcome$by$wickets, " runs"))  }  total\_score=c(total\_score,cummulative\_score)  if(!is.null(jsonData$innings[[inn]]$overs[[over\_index]]$deliveries[[currball]]$wickets)){  cum\_wickets=cum\_wickets+1  wickets=c(wickets,cum\_wickets)  }  else{  wickets=c(wickets,0)  }    currball <- currball + 1  }    balls <- balls + 6  }, error = function(e) {  flag <- FALSE  })  }  }  # print(total\_score)  # print(wickets)  df <- data.frame(  baller = baller\_vector,  batter = batter\_vector,  balls = balls\_vector,  runsperball = runsperball\_vector,  extra = extra\_vector,  total\_score=total\_score,  wickets=wickets,  country = country\_vector,  innings = Innings\_vector,  toss\_decision = toss\_decision,  toss\_winner=match\_winner,  margin=margin,  match\_winner = match\_winner  )  # write.csv(df,"/Users/morampudigopiprashanthraju/Desktop/DataScience/Minor\_Project-II/Data/IND vs AUS/ODI/2003/Conversion\_demo.csv")  # Mushtaq's path  #write.csv(df,"C:\Users\janga\OneDrive\Desktop\Minor\_Project-II/Data/IND vs AUS/ODI/2003/Conversion\_demo.csv")  # print(df)  return( df)  }  # convert\_to\_csv(file\_path) |
| Output | A csv file |
|  |  |
| **Description of Task#2** | **Extraction of runs scored by each player** |
| Required Pre-processing | Generation of another csv |
| Attributes involved | Batter,balls,runs,Bowler,balls |
| R-Code with necessary comments | player\_scores<- function(df){  # Load required library  library(jsonlite)  library(hash)  # Read JSON data  # df <- read.csv("/Users/morampudigopiprashanthraju/Desktop/DataScience/Minor\_Project-II/Data/IND vs AUS/ODI/2003/CSVs/Conversion\_demo.csv")  # df<- read.csv("D:/Minor\_Project-II/Data/IND vs AUS/ODI/2003/CSVs/Conversion\_demo.csv")  # Initialize hash  h <- hash()  h\_bowler <- hash()  prev\_inn=0  for(i in 1:nrow(df)) {  row\_data <- df[i,]    # Update runs for batter  if(!(row\_data$batter %in% names(h))) {    h[[row\_data$batter]] <- list(runs=row\_data$runsperball,out=0,balls=0,country=row\_data$country)  } else {  h[[row\_data$batter]][[1]] <- h[[row\_data$batter]][[1]] + row\_data$runsperball  }  if(!row\_data$extra){  h[[row\_data$batter]][[3]]=h[[row\_data$batter]][[3]]+1  }  # Update runs and wickets for bowler  prev\_score <- 0    if(i != 1 && row\_data$innings==prev\_inn) {  prev\_score <- df[i - 1,]$total\_score  }  prev\_inn=row\_data$innings  if(!(row\_data$baller %in% names(h\_bowler))) {  if(row\_data$country=="Australia"){  h\_bowler[[row\_data$baller]] <- list(score = 0, wickets = 0,extra=0,country="India")  }  else{  h\_bowler[[row\_data$baller]] <- list(score = 0, wickets = 0,extra=0,country="Australia")  }  }  if(row\_data$extra){  h\_bowler[[row\_data$baller]][[3]]=h\_bowler[[row\_data$baller]][[3]]+1  }  if(row\_data$wickets != 0) {  h\_bowler[[row\_data$baller]][[2]] <- h\_bowler[[row\_data$baller]][[2]] + 1  h[[row\_data$batter]][[2]]=1  }    h\_bowler[[row\_data$baller]][[1]] <- h\_bowler[[row\_data$baller]][[1]]+row\_data$total\_score - prev\_score  # print(h\_bowler)  }  h\_df <- as.data.frame(t(sapply(h, unlist)))  h\_df1 <- data.frame(player = rownames(h\_df), h\_df, row.names = NULL)  # write.csv(h\_df, "/Users/morampudigopiprashanthraju/Desktop/DataScience/Minor\_Project-II/Data/IND vs AUS/ODI/2003/players\_details.csv", row.names = FALSE)  # write.csv(h\_df, "/Users/morampudigopiprashanthraju/Desktop/DataScience/Minor\_Project-II/Data/IND vs AUS/ODI/BatsmanScores.csv", row.names = FALSE)  h\_df <- as.data.frame(t(sapply(h\_bowler, unlist)))  h\_df2 <- data.frame(player = rownames(h\_df), h\_df, row.names = NULL)  # write.csv(h\_df, "/Users/morampudigopiprashanthraju/Desktop/DataScience/Minor\_Project-II/Data/IND vs AUS/ODI/2003/players\_details.csv", row.names = FALSE)  # write.csv(h\_df, "/Users/morampudigopiprashanthraju/Desktop/DataScience/Minor\_Project-II/Data/IND vs AUS/ODI/BowlerStats.csv", row.names = FALSE)  return (list(h\_df1,h\_df2))  } |
| Output | CSV file |
|  |  |
| **Description of Task#3** | Extraction of wickets taken by the bowler |
| Required Pre-processing | Extract the performance of a bowler from the ball by ball data |
| Attributes involved | Bowler,balls,runs,wickets |
| R-Code with necessary comments | player\_scores<- function(df){  # Load required library  library(jsonlite)  library(hash)  # Read JSON data  # df <- read.csv("/Users/morampudigopiprashanthraju/Desktop/DataScience/Minor\_Project-II/Data/IND vs AUS/ODI/2003/CSVs/Conversion\_demo.csv")  # df<- read.csv("D:/Minor\_Project-II/Data/IND vs AUS/ODI/2003/CSVs/Conversion\_demo.csv")  # Initialize hash  h <- hash()  h\_bowler <- hash()  prev\_inn=0  for(i in 1:nrow(df)) {  row\_data <- df[i,]    # Update runs for batter  if(!(row\_data$batter %in% names(h))) {    h[[row\_data$batter]] <- list(runs=row\_data$runsperball,out=0,balls=0,country=row\_data$country)  } else {  h[[row\_data$batter]][[1]] <- h[[row\_data$batter]][[1]] + row\_data$runsperball  }  if(!row\_data$extra){  h[[row\_data$batter]][[3]]=h[[row\_data$batter]][[3]]+1  }  # Update runs and wickets for bowler  prev\_score <- 0    if(i != 1 && row\_data$innings==prev\_inn) {  prev\_score <- df[i - 1,]$total\_score  }  prev\_inn=row\_data$innings  if(!(row\_data$baller %in% names(h\_bowler))) {  if(row\_data$country=="Australia"){  h\_bowler[[row\_data$baller]] <- list(score = 0, wickets = 0,extra=0,country="India")  }  else{  h\_bowler[[row\_data$baller]] <- list(score = 0, wickets = 0,extra=0,country="Australia")  }  }  if(row\_data$extra){  h\_bowler[[row\_data$baller]][[3]]=h\_bowler[[row\_data$baller]][[3]]+1  }  if(row\_data$wickets != 0) {  h\_bowler[[row\_data$baller]][[2]] <- h\_bowler[[row\_data$baller]][[2]] + 1  h[[row\_data$batter]][[2]]=1  }    h\_bowler[[row\_data$baller]][[1]] <- h\_bowler[[row\_data$baller]][[1]]+row\_data$total\_score - prev\_score  # print(h\_bowler)  }  h\_df <- as.data.frame(t(sapply(h, unlist)))  h\_df1 <- data.frame(player = rownames(h\_df), h\_df, row.names = NULL)  # write.csv(h\_df, "/Users/morampudigopiprashanthraju/Desktop/DataScience/Minor\_Project-II/Data/IND vs AUS/ODI/2003/players\_details.csv", row.names = FALSE)  # write.csv(h\_df, "/Users/morampudigopiprashanthraju/Desktop/DataScience/Minor\_Project-II/Data/IND vs AUS/ODI/BatsmanScores.csv", row.names = FALSE)  h\_df <- as.data.frame(t(sapply(h\_bowler, unlist)))  h\_df2 <- data.frame(player = rownames(h\_df), h\_df, row.names = NULL)  # write.csv(h\_df, "/Users/morampudigopiprashanthraju/Desktop/DataScience/Minor\_Project-II/Data/IND vs AUS/ODI/2003/players\_details.csv", row.names = FALSE)  # write.csv(h\_df, "/Users/morampudigopiprashanthraju/Desktop/DataScience/Minor\_Project-II/Data/IND vs AUS/ODI/BowlerStats.csv", row.names = FALSE)  return (list(h\_df1,h\_df2))  } |
| Output | Csv file with bowler stats |
|  |  |
| **Description of Task#4** | Extraction of player details through API’s |
| Required Pre-processing | Genertation of all the players details involved in the match |
| Attributes involved | Player,Country,Batting-hand,Bowling-Style |
| R-Code with necessary comments | library(jsonlite)  base\_dir <- "D:/Minor\_Project-II/Data/IND vs AUS/ODI"  year\_folders <- list.files(base\_dir)  #Reading the Aussies's CSV file  players\_details <- read.csv("D:/Minor\_Project-II/Data/IND vs AUS/ODI/players\_details.csv")  for (year in year\_folders){  json\_files <- list.files(paste(base\_dir, year, sep = "/"))  for (json\_file in json\_files){  if (json\_file == "CSVs"){  next()  }  file\_path <- paste(base\_dir, year, json\_file, sep = "/")  print(file\_path)  json\_data <- read\_json(path=file\_path)  aplayers <- json\_data$info$players$Australia  for (player in aplayers){  if (player %in% players\_details$player){  next()  }  else{  players\_details <- rbind(players\_details, data.frame(player=player, country="Australia", batting="None", bowling\_style="None"))  }  }  iplayers <- json\_data$info$players$India  for (player in iplayers){  if (player %in% players\_details$player){  next()  }  else{  players\_details <- rbind(players\_details, data.frame(player=player, country="India", batting="None", bowling\_style="None"))  }  }  }  }  write.csv(players\_details,"D:/Minor\_Project-II/Data/IND vs AUS/ODI/players\_details.csv")  # # write(aus\_players, file = file\_path) |
| Output | A csv file with the details |
|  |  |
| **Description of Task#5** | Plotting of a Kernel Density Plot |
| Required Pre-processing | We had to preprocess the BatsmanScores.csv and BowlerStats.csv and add new fields in like Batting\_Avg, Strike\_Rate and Bowling\_Avg, Eonomy in respective files. |
| Attributes involved | Player, Batting\_Avg, Strike\_Rate for the batsman analysis and Player, Bowling\_Avg, Economy for the bowler analysis |
| R-Code with necessary comments | library(ggplot2)  #For the batsman analysis  data <- read.csv("D:/Minor\_Project-II/Data/IND vs AUS/ODI/BatsmanScores.csv")  aus\_players\_strike\_rate <- data.frame( strike\_rate = numeric())  aus\_players\_batting\_avg <- data.frame( batting\_avg = numeric())  ind\_players\_strike\_rate <- data.frame(strike\_rate = numeric())  ind\_players\_batting\_avg <- data.frame(batting\_avg = numeric())  for (i in 1:nrow(data)) {  if (data$country[i] == "Australia") {  aus\_players\_strike\_rate <- rbind(aus\_players\_strike\_rate, data.frame(strike\_rate = data$strike\_rate[i]))  aus\_players\_batting\_avg <- rbind(aus\_players\_batting\_avg, data.frame(batting\_avg = data$batting\_average[i]))  }  else{  ind\_players\_strike\_rate <- rbind(ind\_players\_strike\_rate, data.frame(strike\_rate = data$strike\_rate[i]))  ind\_players\_batting\_avg <- rbind(ind\_players\_batting\_avg, data.frame(batting\_avg = data$batting\_average[i]))  }  }  aus\_players\_strike\_rate$Group <- "Australia"  ind\_players\_strike\_rate$Group <- "India"  combined\_data <- rbind( aus\_players\_strike\_rate, ind\_players\_strike\_rate)  # Create a KDE plot of strike rates  strike\_rate\_plot<-ggplot(combined\_data, aes(x = strike\_rate, fill = Group, color = Group)) +  geom\_density(alpha = 0.7) +  labs(title = "Kernel Density Estimation Plot of Strike Rates",  x = "Strike Rate",  y = "Density") +  scale\_fill\_manual(values = c("yellow","blue")) +  scale\_color\_manual(values = c("yellow","blue")) +  theme\_minimal()  aus\_players\_batting\_avg$Group <- "Australia"  ind\_players\_batting\_avg$Group <- "India"  combined\_data <- rbind( aus\_players\_batting\_avg, ind\_players\_batting\_avg)  # Create a KDE plot of strike rates  batting\_avg\_plot<-ggplot(combined\_data, aes(x = batting\_avg, fill = Group, color = Group)) +  geom\_density(alpha = 0.7) +  labs(title = "Kernel Density Estimation Plot of Strike Rates",  x = "Batting Average",  y = "Density") +  scale\_fill\_manual(values = c("yellow","blue")) +  scale\_color\_manual(values = c("yellow","blue")) +  theme\_minimal()  library(gridExtra)  grid.arrange(strike\_rate\_plot, batting\_avg\_plot, nrow = 1)  #For the bowler analysis  data <- read.csv("D:/Minor\_Project-II/Data/IND vs AUS/ODI/BowlerStats.csv")  aus\_players\_economy <- data.frame( economy = numeric())  aus\_players\_bowling\_avg <- data.frame( bowling\_avg = numeric())  ind\_players\_economy <- data.frame(economy = numeric())  ind\_players\_bowling\_avg <- data.frame(bowling\_avg = numeric())  for (i in 1:nrow(data)) {  if (data$country[i] == "Australia") {  aus\_players\_economy <- rbind(aus\_players\_economy, data.frame(economy = data$economy[i]))  aus\_players\_bowling\_avg <- rbind(aus\_players\_bowling\_avg, data.frame(bowling\_avg = data$bowling\_average[i]))  }  else{  ind\_players\_economy <- rbind(ind\_players\_economy, data.frame(economy = data$economy[i]))  ind\_players\_bowling\_avg <- rbind(ind\_players\_bowling\_avg, data.frame(bowling\_avg = data$bowling\_average[i]))  }  }  aus\_players\_economy$Group <- "Australia"  ind\_players\_economy$Group <- "India"  combined\_data <- rbind( aus\_players\_economy, ind\_players\_economy)  # Create a KDE plot of strike rates  economy\_plot<-ggplot(combined\_data, aes(x = economy, fill = Group, color = Group)) +  geom\_density(alpha = 0.7) +  labs(title = "Kernel Density Estimation Plot of Economies",  x = "Economy",  y = "Density") +  scale\_fill\_manual(values = c("yellow","blue")) +  scale\_color\_manual(values = c("yellow","blue")) +  theme\_minimal()  aus\_players\_bowling\_avg$Group <- "Australia"  ind\_players\_bowling\_avg$Group <- "India"  combined\_data <- rbind( aus\_players\_bowling\_avg, ind\_players\_bowling\_avg)  # Create a KDE plot of strike rates  bowling\_avg\_plot<-ggplot(combined\_data, aes(x = bowling\_avg, fill = Group, color = Group)) +  geom\_density(alpha = 0.7) +  labs(title = "Kernel Density Estimation Plot of Bowling Averages",  x = "Bowling Average",  y = "Density") +  scale\_fill\_manual(values = c("yellow","blue")) +  scale\_color\_manual(values = c("yellow","blue")) +  theme\_minimal()  library(gridExtra)  grid.arrange(economy\_plot, bowling\_avg\_plot, nrow = 1) |
| Output |  |
|  |  |
| **Description of Task#6** |  |
| Required Pre-processing |  |
| Attributes involved |  |
| R-Code with necessary comments |  |
| Output |  |
|  |  |
| **List of Websites referred** | Cricsheet.com for datasets and many random websites for formulae for calculations in the codes. |
| **Details of Relevant Research papers referred, please present in tabular form the below details for each referred paper (minimum 10 – 12 papers):**   1. Paper Title 2. URL 3. Year of publication 4. Conference / Journal details 5. Dataset used 6. Description of the dataset 7. URL of the dataset 8. Tasks Carried out. 9. Algorithms used for each task 10. Results obtained for each Task as reported in the paper. 11. Gaps reported in the paper |  |

NOTE: The content that you submit can be used for your minor project documentation.