Arlton Gilbert

arltongilbert@gmail.com

Overview

This file will give a greater understanding of how to use this repository. It will cover the files involved and how to run these files.

Process and file explanation

Masters Dissertation 2022

Contents

[Background 2](#_Toc106662078)

[Files Involved 2](#_Toc106662079)

[IPL\_SCHEDULE\_2008\_2020.xlsx 2](#_Toc106662080)

[01 New Seasons.RMD 2](#_Toc106662081)

[IPL\_SCHEDULE\_2008\_2022.xlsx 2](#_Toc106662082)

[Stadiums.xlsx 2](#_Toc106662083)

[02 Player Roles.RMD 2](#_Toc106662084)

[03 WebScraper.RMD 3](#_Toc106662085)

[04 Player Stats.RMD 3](#_Toc106662086)

[05 Player Splitter.RMD 3](#_Toc106662087)

[06 Batsmen Stats.RMD 3](#_Toc106662088)

[07 Bowler Stats.RMD 3](#_Toc106662089)

[08 Stadiums.RMD 3](#_Toc106662090)

[09 Match Grouping.RMD 3](#_Toc106662091)

[10 Modelling.RMD 3](#_Toc106662092)

[11 EDA.RMD 3](#_Toc106662093)

# Background

This is the repository with all the code used in my Masters Dissertation that will hopefully be completed in 2022 currently entitled “Modelling First Innings Totals in T20 Cricket: Applications in the Indian Premier League”. There are many different files that I used in my project and a specific order needed to be followed to run them correctly. I decided that it would be best to create this document which will look at all the files involved and how to run them correctly. If there are any mistakes or suggestions feel free to drop me an email. Now with that out the way, lets jump straight into it

# Files Involved

All programming is done in R using RMD files. These RMD files all create RData files that are then loaded in subsequent files or so that the results can be easily loaded without having to run the whole file again. The files below are in the order they are used.

## IPL\_SCHEDULE\_2008\_2020.xlsx

This is the original dataset which is This comes from <https://www.kaggle.com/narendrageek/can-generate-automatic-commentary-for-ipl-cricket?select=IPL_SCHEDULE_2008_2020.csv> which is then converted to excel and some of the last few rows of data which had missing data was filled in manually.

## 01 New Seasons.RMD

The dataset above was only completed until the end of the 2020 season. Since then both the 2021 and 2022 seasons have completed. We would like to include these seasons in our data as well. This file gets all the relevant information from the last 2 seasons which it then populates in the IPL\_SCHEDULE\_2021\_2022 Extension.xlsx file in the Data folder.

## IPL\_SCHEDULE\_2008\_2022.xlsx

This then combines the IPL\_SCHEDULE\_2008\_2020.xlsx file with the IPL\_SCHEDULE\_2021\_2022\_Extension.xlsx one giving us a full set of fixtures. The other 2 sheets in this file were just used for some checks to see whether our eventual dataset matched the one we would get doing it manually by removing fixtures we didn’t want to use such as abandoned matches. More on that later.

## Stadiums.xlsx

This is where we figure out who is the home team in every fixture. We first pull out the fixtures in the ‘Fixtures’ tab. In the ‘Stadium Count’ we see the number of times each team played in each stadium in non playoff matches. In the ‘Stadiums Assigned’ we record some notes on the usage of stadiums as well as record information on the stadium which we get from the sources listed. On the fixtures tab we then populate the home team using this information. This then gets pulled into the IPL\_SCHEDULE\_2008\_2022.xlsx workbook

## 02 Player Roles.RMD

This reads in the IPL\_SCHEDULE\_2021\_2022\_Extension.xlsx file. This then goes into every match and gets through the team sheets and gets the Role information of every player involved if they are not already in the list. It saves all this roles information in the Player Roles folder under ‘profilelist.xlsx’. Sometimes the info for some players is randomly skipped. You can check the excel file or the profileshort table in the RMD file to make sure there are no empty cells. If it just a few you can just populate them manually in the RMD file and run the rest of the code so that this filters through to the Excel file.

## 03 WebScraper.RMD

This opens the match files one by one and saves them into Excel files with the match information and different innings in different tabs.

## 04 Player Stats.RMD

This file runs through all the scorecards in he Excel files and creates a list of batsmen and bowlers. Review the log.txt file in the Player Stats folder to make sure all matches were processed properly. For the batlist review the batting orders to make sure that it makes sense.

## 05 Player Splitter.RMD

The reads in the ’04 Player Stats.RData’ file and splits them into the different players.

## 06 Batsmen Stats.RMD

This loads the ’02 Player Roles.RData’ and ’05 Player Splitter.RData’ files and calculates the Batsmen Player Strength weightings for each of the matches. Review the log files to ensure that all batsmen were processed properly.

## 07 Bowler Stats.RMD

Like above but for the bowlers.

## 08 Stadiums.RMD

Reads in the ’04 Player Stats.Rdata’ and uses it to create the Median Innings Scores for each of the stadiums.

## 09 Match Grouping.RMD

This loads the ’06 Batsmen Stats.RData’, ’07 Bowler Stats.Rdata’ and ’08 Stadiums.RData’. For each player it adds the player strength ratings that they would have went into the match with.

## 10 Modelling.RMD

This is where the modelling is done.

## 11 EDA.RMD

This loads the ’06 Batsmen Stats.RData’ and ’10 Modelling.Rdata’ files. This is where the Exploratory Data Analysis is done.