Data Analysis For (Cricket) Tragics

"DAFT: When you need to know the game inside out"

Are you looking to analyse cricket data at the next level? Then you need DAFT!

Cricket statistics provide a fascination for people from many countries around the world. They are available from multiple websites, but do you feel you would like to analyse the data at a deeper level? Or are you a professional in the cricket information industry that is looking for a fresh new way to generate unique statistics for your clients? If so, read on. . . .

Data analysts use many different products for this type of work, but we have selected what is one of the best available and the easiest to learn: *RStudio and the Tidyverse.*

If you are not familiar with these products then fear not, they are probably the easiest way for anyone to get into serious data analysis. Even with minimal coding experience.

The people behind the data

Now, the hardest part of data analysis is data curation. This is the gathering and organising of the data. For you, this has been taken care of by Ric Finlay and Dr Jim Palfreyman.

Ric Finlay has devoted a lifetime to recording cricket in the public forum, having commenced a long and productive professional career in 1983, when he first worked for the ABC in Hobart, Tasmania. Since then, he has regularly scored and provided statistical information for ABC Radio, where is the resident cricket analyst, through ABC Grandstand, as well as through commercial television, the ABC Cricket Book and various hard-print media outlets.

Ric is the author of books on Tasmanian representative cricket, and Alan Kippax, and is co-compiler of the popular Tastats computer database, CSW.

Ric has a Bachelor of Arts degree and a Diploma of Education from the University of Tasmania, and in 2008 retired as a teacher of Mathematics at a Hobart Senior Secondary College to concentrate on providing cricket statistics full time. He is a committee member of the Association of Cricket Statisticians and Historians, the world's premier cricket research organisation. He runs a popular cricket Twitter account @RicFinlay.

CSW is a significant database that provides extensive analysis of cricket through its storage of over 68,000 matches, with over 33,000,000 balls bowled, taking over 685000 wickets, and nearly 20,000,000 runs made. This covers all international cricket (Tests, One Day Internationals, and T20 internationals) for both men and women, and comprehensive coverage of Australian, New Zealand, South African and English domestic cricket. The database also takes in all the major T20 leagues around the world, a new force in world cricket. DAFT: Data Analysis For (Cricket) Tragics will give you full access to this mine of information.

Turning Ric's data into something that can be analysed using RStudio and the Tidyverse has been done by Dr Jim Palfreyman. Jim has decades of experience in computing and data analysis including a degree in Mathematics, an Honours degree in Computer Science, a Masters in Astrophysics, and a PhD in Astrophysics. Jim also appears regularly on ABC Radio.

How the data is distributed

The data and code are distributed via GitHub in "Tidy" dataframes (in the form of CSV files) and a file containing R script segments to analyse the data.

This R script file shows you how the data can be analysed and helps you start your own data analysis.

Each are updated regularly. Git and GitHub are the preeminent distribution methods of code and data on the planet. You may never have heard of them, but they are the unsung heroes of software of the whole internet. You can trust them.

Along with the data provided, is sample code to answer some of the quirky questions that have already been asked - e.g. "who played with both Rod Marsh and Shane Warne (but obviously not at the same time)?".

You will also find, for example, the code to generate a unique listing, in rank order, of every cricketer, in both batting or bowling, calculated by comparing each player to the performances of their contemporaries.

Code will be updated as further statistical questions are asked and answered.

RStudio and the Tidyverse

If you are unfamiliar with the Tidyverse, Hadley Wickham's book "R for Data Science" is available free here:

https://r4ds.had.co.nz

If you are new to R, we would advise not bothering to dive in and learn Base R at this stage. Start with the Tidyverse and only learn the bits of Base R that you need.

To get going, you will need to download the latest version of "R", "RStudio", and install the Tidyverse. These are free and run on just about any platform (Linux, Mac, Windows). This page from Hadley's book gives complete instructions:

https://r4ds.had.co.nz/introduction.html#prerequisites

How it works

Once subscribed to the full version of DAFT, you get access to the DAFT GitHub repository which gives you access to already created code and worldwide cricket data that is updated regularly. (We strive for daily updates, but sometimes life gets in the way.)

The data is provided in CSV (comma separated value) format files, which is a universal form of providing data. It is not ideal, but for a straightforward cricket dataset this works well. You can even import this into a spreadsheet if you wish.

However, we provide R code (using RStudio) to load the data in and use for analysis. We highly recommend you use this method.